Lecture Notes on Mathematical Systems Biology

[*Continuously revised and updated. This is version 6.7. Compiled April 30, 2015*]

Please address comments, suggestions, and corrections to the author.

Eduardo D. Sontag, Rutgers University, ©2005,2006,2009,2010,2011,2013,2014

The first version of these notes dates back to 2005. They were originally heavily inspired by Leah Keshet's beautiful book *Mathematical Models in Biology* (McGraw-Hill, 1988), and the reader will notice much material "borrowed" from there as well as other sources. In time, I changed the emphasis to be heavier on "systems biology" ideas and lighter on traditional population dynamics and ecology. (Topics like Lotka-Volterra predator-prey models are only done as problems, the assumption being that they have been covered as examples in a previous ODE course.) The goal was to provide students with an overview of the field. With more time, one would include much other material, such as Turing pattern-formation and detailed tissue modeling. Starting with Version 6, there is a first very short chapter on difference equations, which can be skipped without loss of continuity. (And, conversely, can be covered by itself.)

The writing is not always textbook-like, but is sometimes "telegraphic" and streamlined, so as to make for easy reading and review. (The style is, however, not consistent, as the notes have been written over a long period.) Furthermore, I do not use "definition/theorem" rigorous mathematical style, so as to be more "user-friendly" to non-mathematicians. However, the reader can rest assured that every statement made can be cast as a theorem! Also, I tried to focus on intuitive and basic ideas, as opposed to going deeper into the beautiful theory that exists on ordinary and partial differential equation models in biology – for which many references exist.

Please note that many figures are scanned from books or downloaded from the web, and their copyright belongs to the respective authors, so please do not reproduce.

Originally, the deterministic chapters (ODE and PDE) of these notes were prepared for the Rutgers course Math 336, Dynamical Models in Biology, which is a junior-level course designed for Biomathematics undergraduate majors, and attended as well by math, computer science, genetics, biomedical engineering, and other students. Math 336 does not cover discrete methods (genetics, DNA sequencing, protein alignment, etc.), which are the subject of a companion course. With time, the notes were extended to include the chapter on stochastic kinetics, covered in Math 613, Mathematical Foundations of Systems Biology, a graduate course that also has an interdisciplinary audience. In its current version, the material no longer fits in a 1-semester course. Without the stochastic kinetics chapter, it should fit in one semester, though in practice, given time devoted to exam reviews, working out of homework problems, quizzes, etc., this is unrealistic.

Pre-requisites for the deterministic part of notes are a solid foundation in calculus, up to and including sophomore ordinary differential equations, plus an introductory linear algebra course. Students should be familiar with basic qualitative ideas (phase line, phase plane) as well as simple methods such as separation of variables for scalar ODE's. However, it may be possible to use these notes without the ODE and linear algebra prerequisites, provided that the student does some additional reading. (An appendix provides a quick introduction to ODE's.) The stochastic part requires good familiarity with basic probability theory.

I am routinely asked if it is OK to use these notes in courses at other universities. The answer is, obviously, "of course!". I do strongly suggest that a link to the my website be provided, so that students can access the current version. And please provide feedback!

Acknowledgements

Contents

Chapter 1

Difference Equations

In this short chapter, we provide a brief introduction to scalar difference equations.[1] With differential equations, things get interesting only in higher dimensions. On the other hand, with difference equations, certain behaviors, such as periodic orbits and chaos, can already appear in models with just one variable. In addition, it is possible to understand this material without knowledge of calculus nor linear algebra, and as such, it provides an easy introduction to ideas about dynamics.

1.1 Iterations, the P^+ and $\triangle$ formalisms, and exponential growth

We use t to denote the time variable, which in this chapter is taken to be discrete, $t = 0, 1, 2, \ldots$.

It is implicitly assumed that we have fixed a unit of time measurement; for example, t may be measured in generations if we are interested in genetics, or in hours, days, or years if studying population size for cells, flies, or humans respectively.

We use the letter P for a function that describes a time-dependent quantity that we wish to study, such as the size of a certain population. Depending on the context, we find sometimes more convenient to write "P_t" instead of $P(t)$" for the population at time t.

A difference equation is a just rule that tells us how the population at the next time $t + 1$ depends on the population at the current time t:

$$P_{t+1} = F(P_t)$$

where "P" is some scalar function. It is usually too hard to find closed-form solutions of such equations, but numerical experiments can help tremendously, especially if coupled with qualitative understanding of the type that we describe next.

It is sometimes more convenient (and it provides better intuition about the transition to differential equations) to think of a rule that tells us not what the next state will be, but instead quantifies the *change* in population that will be observed. We write this increment as "$\triangle P_t$" defined mathematically as follows:

$$\triangle P_t := P_{t+1} - P_t \,.$$

---

[1] The material here is largely "borrowed" from part of Chapter 1 of *Mathematical Models in Biology: An Introduction*, 2004, by Allman and Rhodes. The reader is encouraged to consult that text for much more material, including vector systems of difference equations and for applications to molecular evolution, phylogenetic trees, classical genetics, and epidemics.

Of course, it is obvious from this definition that we can compute P_{t+1} by starting from P_t and adding the change to it:

$$P_{t+1} = P_t + \Delta P_t \,.$$

In other words, a rule for specifying ΔP_t, or a rule for specifying P_{t+1}, as functions of P_t are just different ways of conveying the same information.

We adopt the following convention, to be used whenever t is clear from the context. Instead of $P_{t+1} = F(P_t)$, we write

$$P^+ = F(P)$$

(think of the superscript "+" as "step the index up by one") and instead of ΔP_t we write ΔP, dropping t.

1.1.1 Linear equations

Let us start with a very simple example, basically the exponential growth model described by Malthus in the late 18th century (which may be treated also using differential equations, if time is assumed continuous instead of discrete). Suppose that at each time t, the population *change* is as follows:

1. add a multiple f of the population (think of a "fecundity" due to births), and

2. subtract a fraction d of the population (think of a death rate).

Given how we said this, this specifies a rule for the change ΔP, namely:

$$\Delta P_t = f P_t - d P_t = (f - d) P_t \,.$$

Sometimes we write this in short form, as:

$$\Delta P = (f - d) P$$

but we should always remember the arguments t in both sides. Of course we also have a rule for computing the next P^+:

$$P^+ = P + \Delta P = P + (f - d)P = (1 + f - d)P = \lambda P \,,$$

where we write

$$\lambda := 1 + f - d$$

for convenience (because only the value of λ, and not the individual fecundity and death rate, matter for the subsequent analysis); λ is called the growth rate of the population.

As a trivial example, suppose that $f = 0.1$, $d = 0.02$, so that $\lambda = 1 + f - d = 1.08$. If we specify an "initial condition" such as $P_0 = 100$, we may recursively compute:

$$P^+ = 1.08P, \quad P_0 = 100$$

by repeatedly multiplying by 1.08. We get

Day	Population
0	100
1	$(1.08)100 = 108$
2	$(1.08)^2 100 = 116.64$
3	$(1.08)^3 100 \approx 125.971$
4	$(1.08)^4 100 \approx 136.0489$

and, obviously, we know the solution for all t in closed form: $P = 100(1.08)^t$.

What is the meaning of non-integer populations? It depends on the units in which P is being specified. For example, P_t may be measured in millions of individuals (cells, people) at time t, in which case fractions have an obvious meaning: 136.0489 means 1,360,489 individuals. Or, we may be measuring populations by weight, such as in tons if dealing with a harvest of a crop or fish. In any event, all models are ultimately simplifications of reality, so we may view a number like "136.0489" as simply "approximately 136" and ignore the fractional part.

A word description of such a problem might be as follows. Suppose we study an insect species for which, in each generation, each female lays 150 eggs.[2] We also assume that when eggs hatch, only 2% survive to become adult females. (The remaining eggs are assumed to not hatch or to be males.) We assume that we measure time in generations, and all adults die before the next generation size is computed. Thus, the death rate is $d = 1$ (everyone dies) and the effective fecundity is

$$f = .02(150) = 3$$

which leads to the following equation for the female population:

$$P^+ = (1 + 3 - 1)P = 3P$$

(we ignored males in order to simplify the model; assume that enough males are available that the model makes sense).

---

[2]Obviously, this is nonsense. Not every female will lays the exact same number of eggs. To keep things simple, however, we assume that they are all perfect and lay precisely 150 eggs. In probabilistic terms, we are looking at mean or expected values.

1.2   Some nonlinear models

The exponential growth model is, of course, not realistic: in practice, birth and death rates are dependent on available resources (such as food, water, or space), and thus, at least indirectly, on the size of the population. Thus, it is reasonable to modify the linear growth model by assuming that the *growth rate decreases as the population size increases*. This is sometimes called the "density-dependent growth" model.

To introduce the model, let us consider the per-capita growth rate over a single time step, that is to say, the change in population per individual $\Delta P_t / P_t$ that happens between times t and $t + 1$.

In the usual exponential growth model, we have a constant $\Delta P_t / P_t \equiv \lambda$.

It is reasonable to assume that, for small populations, there are sufficient environmental resources to support a positive per capita growth at rate r, but that for large populations the per-capita growth is smaller as individuals compete for both food and space. For even larger populations, say, larger than some number K (called the "carrying capacity" of the environment), the per-capita growth rate should be negative, as there are insufficient resources to maintain the population size (some will starve and will not be able to reproduce; more will die).

So, we want a function that expresses that $\frac{\Delta P}{P}$ decreases with increasing P, eventually becoming negative, as shown in the left panel in the figure.
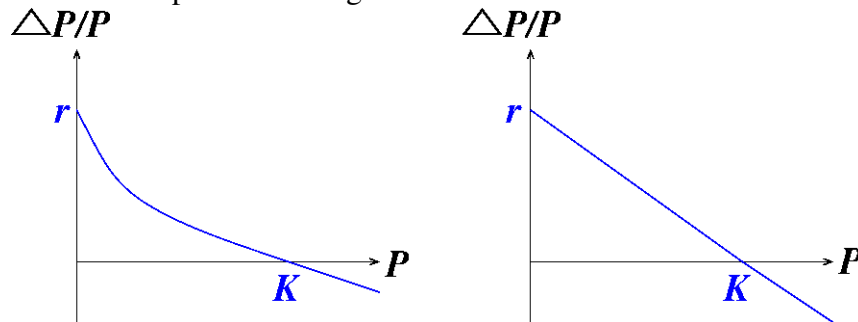
The simplest such function is a linear function as shown in the right panel:

$$\frac{\Delta P}{P} \;=\; r\left(1 - \frac{P}{K}\right)\,.$$

In other words, $\Delta P_t = r P_t \left(1 - \frac{P_t}{K}\right)$, and, since $P_{t+1} = P_t + \Delta P_t$, the model can also be described by:

$$P^+ \;=\; P\left(1 + r\left(1 - \frac{P}{K}\right)\right)$$

One calls this the *logistic model*. (To be more precise, one should call it the "discrete" logistic model, to differentiate it from the continuous model studied for ODE's.) Note that when the population is small (meaning that $P \ll K$), $P/K \approx 0$ so we have

$$P^+ \approx \lambda P\,, \quad \text{where } \lambda = 1 + r$$

– in other words, when the population is far below the carrying capacity, we have a behavior just as in the simpler exponential growth model.

There is no simple expression for the iterates, comparable to "$\lambda^t P_0$" for the exponential growth model. However, we can numerically iterate in order to have some feel for the solutions. It is rather amazing

that this simple model can lead to very interesting and unexpected behaviors, as we see later. For some parameters, however, the behavior is not too surprising. For example, if we iterate

$$P^+ = P(1 + .7(1 - P/10)), \qquad P_0 = 0.4346$$

we obtain:

$P_1 = .7256$, $P_2 = 1.1967$, $P_3 = 1.9341$, $P_4 = 3.0262$, $P_5 = 4.5034$, $P_6 = 6.2362$,

$P_7 = 7.8792$, $P_8 = 9.0489$, $P_9 = 9.6514$, $P_{10} = 9.8869$, $P_{11} = 9.9652$, $P_{12} = 9.9895$, ...

as plotted below.[3]
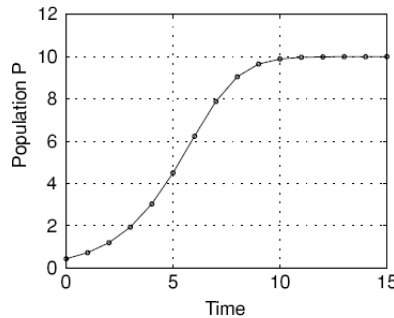
(Only values at integer times t are computed; the interpolating line segments are only shown to help your eyes to follow the behavior over time.)

Note that the population increases monotonically toward the carrying capacity value of 10, first slowly, then more rapidly, and finally slowing up again. A "sigmoidal" picture, as often seen in lab experiments, is observed.

1.2.1 Cobwebbing

A most useful qualitative tool for understanding nonlinear discrete iterations is as follows. Consider the same example as earlier:

$$P^+ = P \left(1 + .7 \left(1 - P/10\right)\right), \quad P_0 = 2.3$$

and now proceed as follows:

1. graph the parabola (defined by the equation that specifies P^+ in terms of P);

2. graph the diagonal line $P^+ = P$;

3. mark the point $(P_0, P_0) = (2.3, 2.3)$ on the diagonal;

4. to find P_1, move *vertically* toward to the graph of the parabola, to reach the point (P_0, P_1);

5. to find P_2, we first need to mark P_1 on the x-axis, or equivalently we mark (P_1, P_1) on the diagonal; we do this as follows: move *horizontally* from (P_0, P_1) toward the diagonal, hitting it at (P_1, P_1) (thus, we kept the y-second coordinate same, and changed the x-coordinate);

6. finally, to find P_2, we move vertically back to the parabola to find (P_1, P_2);

7. now iterate the procedure "vertically to parabola, horizontally to diagonal, vertically to parabola" until the pattern becomes obvious.

---

[3]Most figures in this chapter are reproduced, or generated using their MATLAB scripts, from Allman and Rhodes's book.

The figure shown below illustrates the procedure.
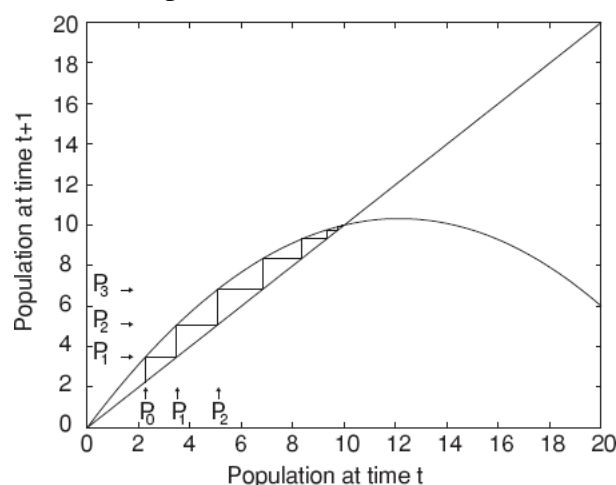
It is clear from the graph that, whenever the initial population P_0 is between 0 and $K=10$, the population increases, asymptotically approaching the carrying capacity K (formally, $\lim_{t\to\infty} P_t = K$).

Still with the same parameters $r=0.7$ and $K=10$, let us now start from $P_0 = 18$. With this P_0, and more generally whenever $P_0 > K=10$, the population also has $\lim_{t\to\infty} P_t = K$. The approach back to K might be monotonically decreasing, or, if $P_{t+1} < K$, there is first an "undershoot" followed by a recovery.
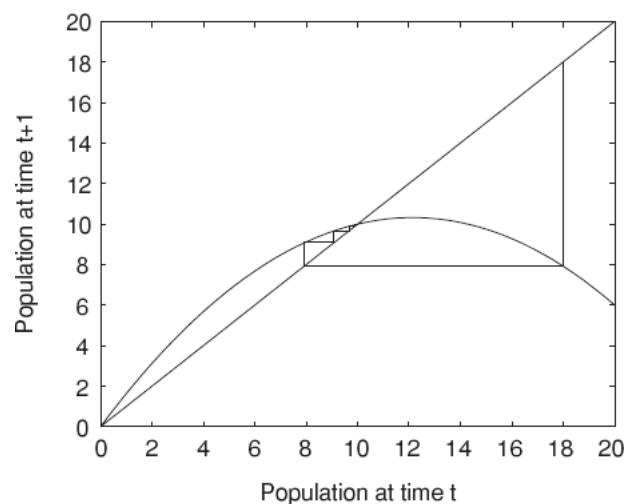
Actually, when the population starts extremely high, the population could become negative, which is of course nonsense. Therefore, the model is still very unrealistic. One way to fix the model is to just truncate the parabola at zero, indicating extinction of the population, but many other changes are possible.[4]

Incidentally, the MATLAB code used to plot the parabola and the diagonal is as follows:

```
x=0:0.01:20
y=x.*(1+0.7*(1-x));
plot(x,y,x,x,'linewidth',2)
```

Note the use of ".*" because we want a component-wise product. We are using "linewidth" of 2 to get thicker pictures for printing.

---

[4]See Section 1.4 in Allman-Rhodes for many proposed models.

1.3 Equilibria and linearizations

In the logistic example, with the parameters used earlier ($r = 0.7$ and $K = 10$), one can verify that $P_t \to K = 10$ as $t \to \infty$, no matter what is the initial state, with the only exception of the very special state in which the population is empty, $P_0 = 0$. Note that, in particular, if $P_t = 10$ exactly, then $P_{t+1} = 10$. Each of 0 and 10 is a steady state.

More generally, a *steady state* of $P^+ = F(P)$ (also called an *equilibrium* or a *fixed point*) is a value P^* with the property that

$$F(P^*) = P^*$$

or equivalently, if we write the iteration as $\Delta P = G(P)$, a steady state must satisfy that there is no change:

$$G(P^*) = 0$$

(this second condition makes it easier to compare to equilibria in ODE's).

Graphically, equilibria correspond to the intersection of the graph of $F(P)$ with the diagonal, and algebraically, they are obtained by solving $P = F(P)$. For example, solving the quadratic equation $P = P(1 + .7(1 - P/10))$ for P gives the two solutions $P = 0$ and $P = 10$.

Although they are both equilibria, there is a major difference between 0 and 10. A population that starts near 0 tends to move away from 0, but a population that starts near 10 tends to move toward 10.
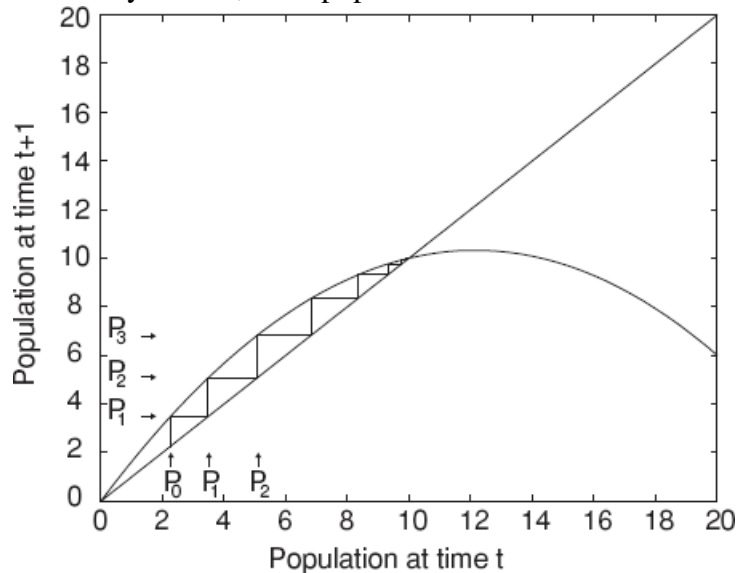
Mathematically, we say that $P^* = 0$ is an *unstable* or *repelling* equilibrium while $P^* = 10$ is a *stable* or *attracting* equilibrium.

To be precise, an equilibrium value P^* is said to be (locally, asymptotically) *stable* if the following property holds:

For each $\varepsilon > 0$ there is some $\delta > 0$ such that

$$|P_0 - P^*| < \delta \quad \Rightarrow \quad |P_t - P^*| < \varepsilon \;\; \forall\, t > 0 \quad and \quad \lim_{t \to \infty} P_t = P^*$$

This definition says that provided that we start near P^*, we will never deviate too much from P^* (no "large excursions"), and eventually will asymptotically approach it.

For the purposes of these notes, we'll just say "stable" and understand the property as saying that $\lim_{t\to\infty} P_t = P^*$ provided that we start from a P_0 near enough P^*.

Note that the notion of stability is "local" in the sense that we only ask about trajectories that start "close enough" to P^*. Certainly, if there is another equilibrium (such as "0" in our previous example), then starting from that other equilibrium, there is no way that we'll ever approach P^*, even if P^* is stable!

On the other hand, in practice it is very difficult to ever see a real system at one of its unstable equilibria, because the smallest perturbation will take us away from that state. Think of a pen perfectly balanced in a vertical position, of a ball placed at the top of a hill. Thus, stable equilibria are of great interest.

1.3.1   Linearization at an equilibrium P^*

As stability depends on what happens close to an equilibrium, we look at the deviation from an equilibrium P^*:

$$p_t := P_t - P^*$$

and, since

$$|p_{t+1}| < |p_t| \quad \Rightarrow \quad P^+ \text{ moves closer to } P^*$$
$$|p_{t+1}| > |p_t| \quad \Rightarrow \quad P^+ \text{ moves away from } P^*$$

all we need to do is to understand if the ratio $\left|\frac{p_{t+1}}{p_t}\right|$ is < 1 or > 1.

We use calculus to figure this out, *remembering that the perturbation p_t is ≈ 0.*

Since

$$p_{t+1} \;=\; P_{t+1} - P^* \;=\; F(P_t) - P^* \;=\; F(P^* + p_t) - P^* \,,$$

we have:

$$\frac{p_{t+1}}{p_t} \;=\; \frac{F(P^* + p_t) - F(P^*)}{p_t} \;\approx\; F'(P^*)$$

where we used the definition of derivative, which says that $\frac{F(x+h)-F(x)}{h} \to F'(x)$ as $h \to 0$. In other words,

$$p_{t+1} \;\approx\; F'(P^*)p_t \,.$$

With a little more formalism, one can in fact prove rigorously that if P^* is an equilibrium for $P^+ = F(P_t)$, then:

$$|F'(P^*)| < 1 \quad \Rightarrow \quad P^* \text{ stable}$$
$$|F'(P^*)| > 1 \quad \Rightarrow \quad P^* \text{ unstable} \,.$$

One often calls $F'(P^*)$ the linearization at the given equilibrium.

Let us revisit our previous example $F(P) = P(1 + .7(1 - P/10))$ using linearizations. We know that the equilibria are 0 and 10. Now, in general, $F'(P) = (1 + .7(1 - P/10)) + P(.7)(-1/10)$, so, in particular: $F'(0) = 1.7 > 1$, confirming that 0 is unstable, and $F'(10) = 1 - .7 = 0.3 < 1$, confirming that 10 is stable.

When $|F'(P^*)| = 1$, one cannot decide using just first derivatives.[5] To illustrate this, take the example $F(P) = P + P^2$ and the equilibrium $P^* = 0$. Note that $F'(P^*) = 1$. One can see graphically (using cobwebbing) that starting at $P_0 = -0.01$ results in convergence to 0, but starting at $P_0 = 0.01$ results in divergence from 0, so this state is in fact neither stable nor unstable.
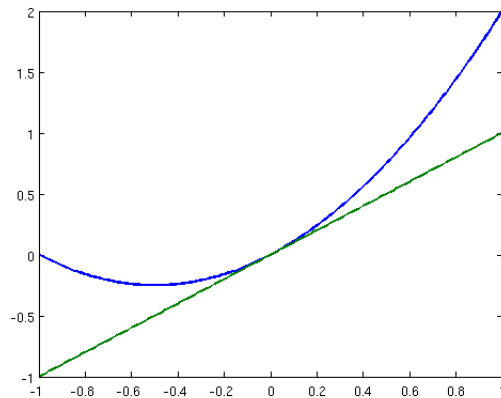
graph of $F(P) = P + P^2$ and diagonal

As an optional homework problem, you may want to analyze these two other cases where $F'(0) = 1$: $F(P) = P + P^3$ and $F(P) = P - P^3$.

1.4 Oscillations, Bifurcations, and Chaos

Let us continue the analysis of the logistic model $P^+ = P(1 + r(1 - P/K))$. For simplicity, we'll use only $K = 1$ from now on:

$$F(P) = P(1 + r(1 - P)) = P(1 + r - rP)$$

but we will investigate the behavior of this system for different positive r's (before, we had studied the special case $r = 0.7$).

No matter the value of r, there are only two equilibria, $P^* = 0$ and $P^* = 1$, and we can easily compute the linearizations at each: $F'(0) = 1 + r$, $F'(1) = 1 - r$. Thus, 0 unstable (for any given value of the parameter $r > 0$), but what about $P^* = 1$? Now things get really interesting! We study various examples one at a time.

1.4.1 $0 < r \leq 1$

Since $F'(1) = 1 - r$ and $0 \leq 1 - r < 1$, $P^* = 1$ is stable. Moreover,

$$p_{t+1} \approx F'(1)p_t = (1 - r)p_t$$

shows that the sign of p_t does not change. Therefore, not only does the perturbation shrink, but an initially positive perturbation remains positive and an initially negative one remains negative. In other words, the population moves toward the equilibrium $P = 1$ without overshoot (assuming $P_0 \approx P^*$). You should see this by performing a cobwebbing on the graph shown for the example $r = 0.5$.

---

[5]This is quite analogous to, when checking for local minima and maxima, we get a zero second derivative - we may have a min, a max, or an inflection point.

1.4.2 $1 < r < 2$

As $F'(1) = 1-r$ and $-1 < 1-r < 0$, the equilibrium $P^* = 1$ is still stable, but since $p_{t+1} \approx (1-r)p_t$, the sign of p_t alternates between positive and negative as t increases. In other words, we expect to see an oscillatory behavior above and below the equilibrium, as the perturbation from equilibrium alternates in sign. *The population approaches equilibrium as a damped oscillation.* You should see this by performing a cobwebbing on the graph shown for the example $r = 1.5$.
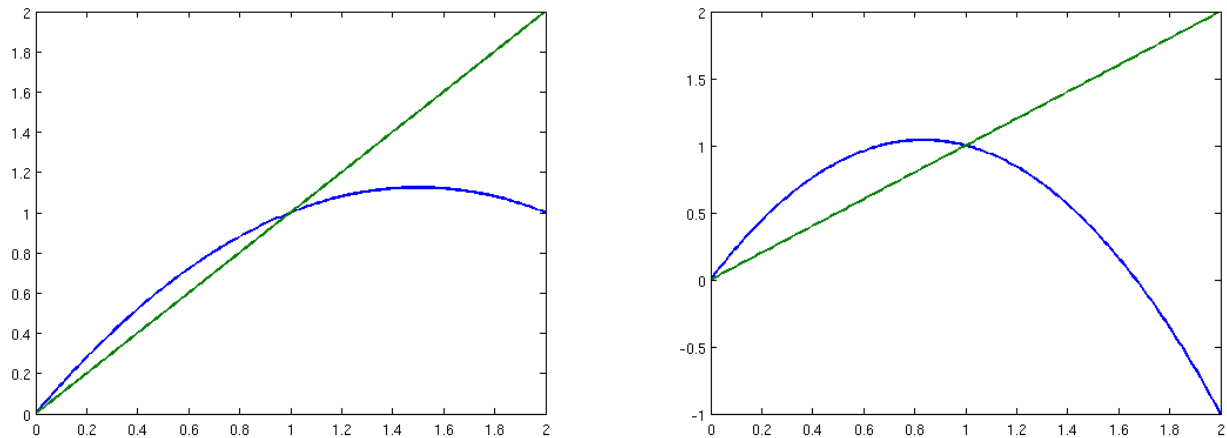
$F(P)$ and diagonal when $r = 0.5$ and $r = 1.5$

1.4.3 $r > 2$

Now $F'(1) = 1 - r$ and $1 - r < -1$, so $P^* = 1$ is unstable.

We first study the special case, $r = 2.4$. It is hard to see exactly what happens here. So we first plot a few iterates (with an initial condition of 0.3):
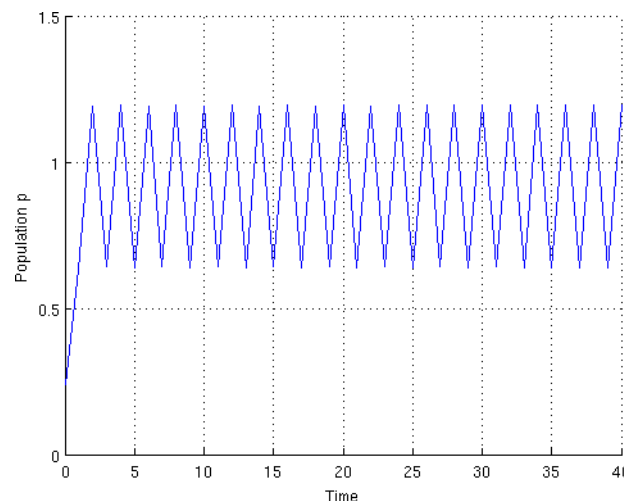
Some iterates when $r = 2.4$, quickly approaching periodic orbit

To algebraically find this period-2 oscillation, that is to say a point P_0 such that $P_1 = F(P_0)$ and $P_2 = F(P_1) = P_0$, we write

$$F^2(P) = F(F(P)) = F(P)(3.4 - 2.4F(P)) = P(3.4 - 2.4P)(3.4 - 2.4P(3.4 - 2.4P))$$

and solve $F^2(P) = P$ for a fixed point P. This gives 0 and 1 (which also solve $F(P) = P$, so we already knew about that), as well as two new solutions: $\approx 0.6402812675, 1.193052066$.

Cobwebbing confirms a period-2 oscillation when starting from one of these states:
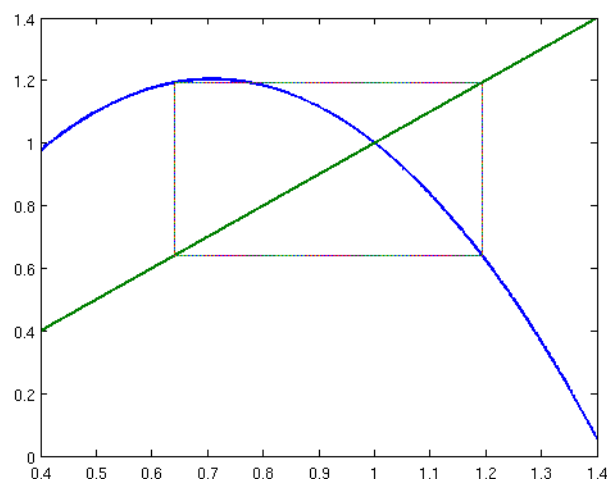
Cobwebbing for $r = 2.4$

As r is increased further, the values in the 2-cycle change, but the existence of some 2-cycle persists, until we hit another value of r, where a new qualitative change occurs this time we see the 2-cycle becoming a 4-cycle.

For example, let us study another special case, $r = 2.5$ and again start by plotting a few iterates, discovering that there is a period-4 orbit:
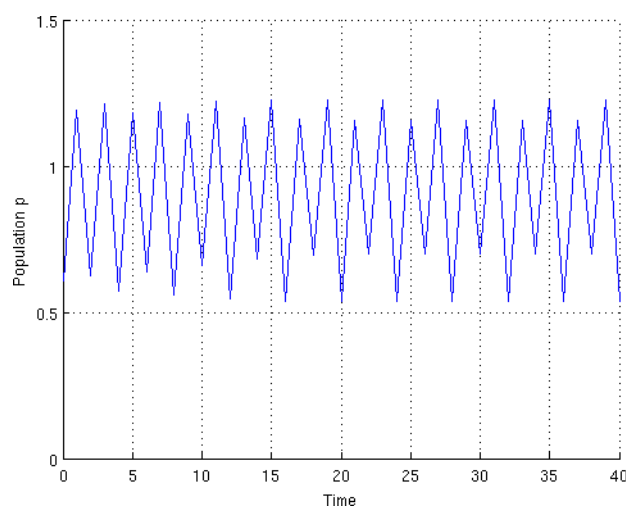
To 4 decimal digits, the numbers of this orbit are numerically found to be: 1.2250, 0.5359, 1.1577, 0.7012.

Further increases in r produce an 8-cycle, then a 16-cycle, and so forth. This is an example of what is called the "period doubling route to chaos".

1.4.4   Bifurcations and chaos

To visualize the effect of changing r, we may draw a *bifurcation diagram*, which is produced as follows: For each value of r on the horizontal axis, pick some value P_0, then first iterate "enough times" that transient behavior is over, discarding these values of P_t. We then plot values of consecutive P_t, on a vertical axis above this r.
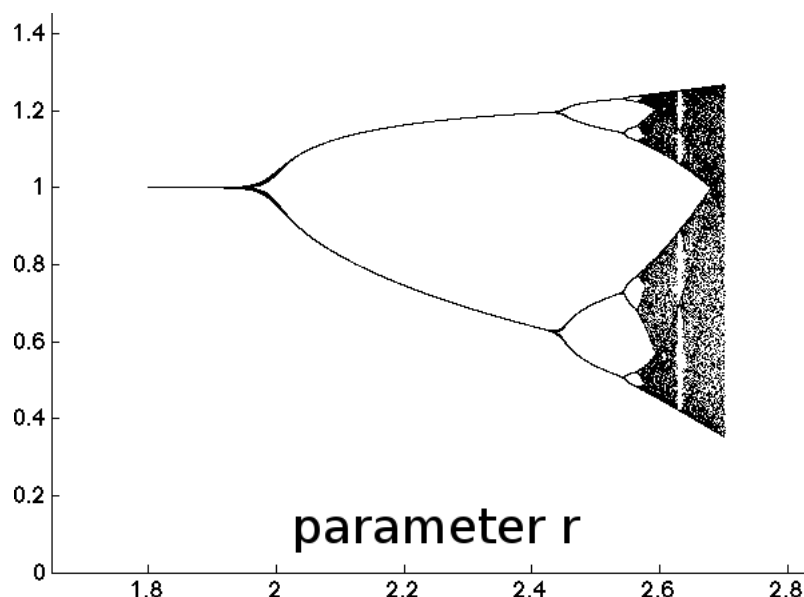
We see that, for $r = 2.4$, there are two values of P_t that appear (after a transient behavior that we disregarded), consistently with our having found a periodic orbit for that value of r. We also see four points when $r = 2.5$, and so forth.

Even more interestingly, when r is increased past a certain point ($\approx 2.692\dots$) all the bifurcations into 2^n-cycles have already taken place, and a new type of behavior emerges: the values seem to be spread out. Take for example $r = 2.75$. A plot of the first few P_t's shows what looks like "random" behavior.
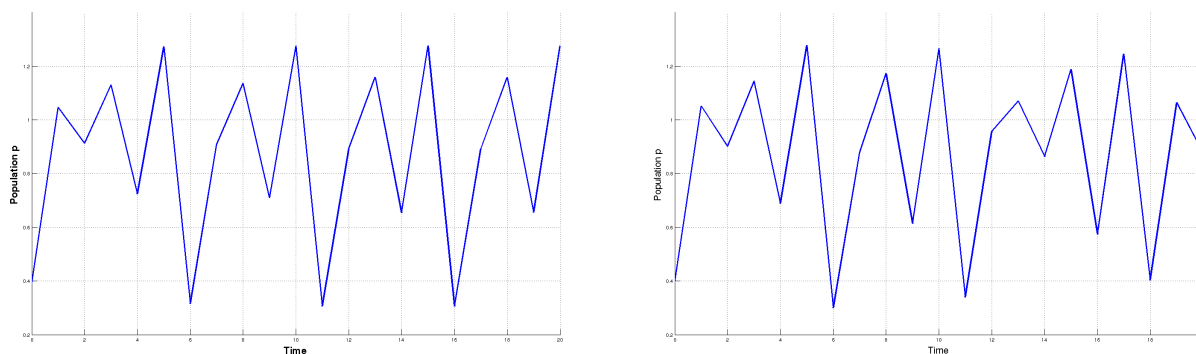
$r = 2.75$, two different initial conditions

Of course there is nothing random (in the sense of coin-flipping): a deterministic formula produces the values. But such irregular behavior is usually referred to as "chaotic". An interesting feature is that of high sensitivity to initial conditions or "butterfly effect" as it was called by the American mathematician and meteorologist Edward Lorenz in 1961.[6] For the two only slightly different initial

---

[6]The butterfly effect is the sensitive dependence on initial conditions, in which a small change in one state of a deterministic nonlinear system can result in large differences in a later state. Lorenz used this term as a metaphor, referring

values shown earlier, the populations change similarly only for a few time steps, but quickly become very different, as clearly seen when superimposed.

_____
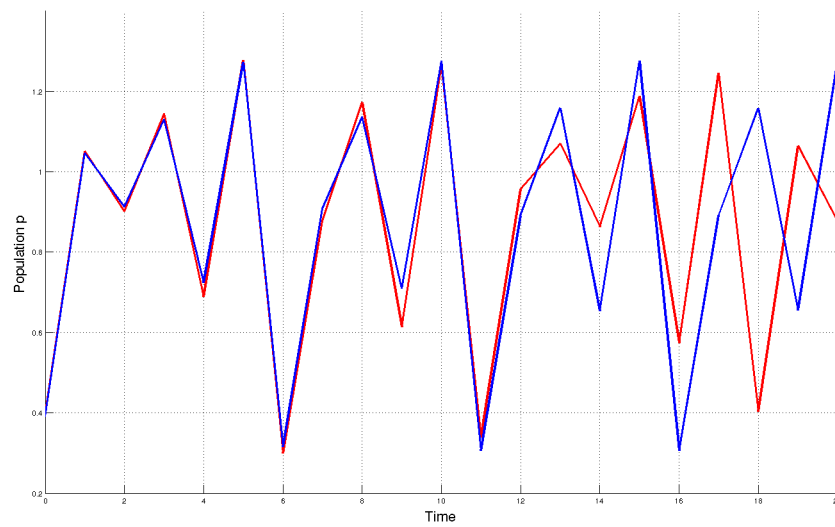
to the details of a hurricane being influenced by minor perturbations such as the flapping of the wings of a distant butterfly several weeks earlier.

1.5 Problems for Scalar Difference Equations chapter

Problems SDE1: Formalisms and exponential growth

1. It is known that a given population triples each hour, due to the net effect of fecundity and deaths.

   (a) Assuming an initial population at time $t = 0$ of 100, compute the population sizes for $t = 1, 2, 3, 4, 5$.

   (b) Show the equations that model the population, in each of the two formalisms:

   (i) provide a formula for P^+ in terms of P

   (ii) provide a formula for ΔP in terms of P.

   (c) What, if anything, can you say about the fecundity and death rates for this population?

2. Suppose that you observe a frog embryo in which all cells divide roughly every half hour. In other works, the number of cells in this embryo doubles every half hour. We start with one cell at time t.

   (a) Write down an equation for P^+ in terms of P that models this situation. Explain what time unit you are using, and what is the initial value P_0.

   (b) How many cells are there after 5 hours?

3. Using a calculator, compute the populations at times $t = 0$ to 6 for the following models:

   (a) $P^+ = 1.3P$, $P_0 = 1$

   (b) $P^+ = .8P$, $P_0 = 10$

   (c) $\Delta P = .2P$, $P_0 = 10$

4. (a) Obtain 20 iterations of $P^+ = 1.3P$, $P_0 = 1$ using MATLAB. You may want to enter the following command sequence:

   ```
   p=1
   x=1
   for i=1:20
     p=1.3*p
     x=[x p]
   end
   ```

   (b) Next, graph your data, using the command:

   ```
   plot([0:20],x)
   ```

   Important note: instead of entering these commands one by one, it is far better to create a file, let us say called "myprogram.m" which contains the above commands. Then you can just type "myprogram" into MATLAB. Make sure to use a basic text editor (such as Notepad, or the MATLAB editor), which does not insert any formatting characters.

5. Fill-in:

   (a) The model $P^+ = kP$ represents a growing population if k is any number in the range

   ☐

   (b) The model $\Delta P = rP$ represents a growing population if r is any number in the range

   ☐

   (c) The model $P^+ = kP$ represents a declining population if k is any number in the range

   ☐

   (d) The model $\Delta P = rP$ represents a declining population if r is any number in the range

   ☐

6. Explain why the model $\Delta P = rP$ cannot be biologically meaningful for describing a population, if r is a number < -1.

7. Suppose that the size of a certain population is affected not only by birth and death, but also by immigration and emigration, and each of these occurs in a yearly amount that is proportional to the size of a population.

   That is, if the population is P, then within a time period of 1 year, the number of births is bP, the number of deaths is dP, the number of immigrants is iP, and the number of emigrants is eP, for some positive constants b, d, i, e.

   Model the population growth by a formula like "$P^+ = \lambda P$" specifying a formula for r in terms of b, d, i, e.

Problems SDE2: Logistic model and basic cobwebbing

1. Using any software of your choice, and testing several different values for P_0, investigate the long-term behavior of the model $\Delta P = rP(1 - P/10)$ for these different parameter values: $r = .2, .8, 1.3, 2.2, 2.5, 2.9$, and 3.1.

   (You may have to vary the number of time steps that you run the model to study some of these)

2. (a) Rewrite the model $P^+ = P + .2P(10 - P)$ in each of the following forms:

   (i) $\Delta P = kP(1 - P/K)$

   (ii) $\Delta P = kP - hP^2$

   (iii) $\Delta P = kP(K - P)$

   (iv) $P^+ = kP - hP^2$

   (pick in each case an appropriate value for the constants k and h).

   (b) Repeat (a) for $P^+ = 2.5P - .2P^2$.

3. Consider the model $\Delta P = .8P(1 - P/10)$.

   (a) Plot ΔP as a function of P. You may want to use, for example, the following MATLAB commands:

   ```
   x=[0:.1:12];
   y=.8*x.*(1-x/10);
   plot(x,y)
   ```

   (b) Construct a table of values of P_t for $t = 0, 1, 2, 3, 4, 5$ starting from $P_0 = 1$.

   (c) Graph P^+ as a function of P.

   (d) On your graph from part (b), construct a cobweb beginning at $P_0 = 1$.

   (You can add the diagonal line $y = x$ to your graph by entering the commands "hold on" and "plot(x,y,x,x)".)

   Compare the values from your cobweb to those that you obtained in part (b).

4. These are measurements of populations obtained in a laboratory experiment using insects:

   $P_0 = .97, P_1 = 1.52, P_2 = 2.31, P_3 = 3.36, P_4 = 4.63, P_5 = 5.94, P_6 = 7.04, P_7 = 7.76, P_8 = 8.13, P_9 = 8.3, P_{10} = 8.36,$

   (a) Plot these points, to convince yourself that they are (roughly) consistent with a logistic model. You may use these commands:

   ```
   times=[0 1 2 3 4 5 6 7 8 9 10];
   Ps=[0.97 1.52 2.31 3.36 4.63 5.94 7.04 7.76 8.13 8.3 8.36];
   plot(times,Ps)
   ```

   (b) Now estimate the parameters r and K in "$\Delta P = rP(1 - P/K)$" by performing the following steps (there are much better methods for estimating sigmoidal functions; we are just being intuitive here):

(i) first give a guess of K by looking at the graph (remember that K is the carrying capacity!);

(ii) then, using that $P_0 \ll K$, approximate $\Delta P_0 / P_0$ by r; what value of r do you obtain?

Plot the data and the iteration, starting at the same P_0, with the values that you obtained.

Finally, use some trial and error, increasing r, to see if you get a better fit.

5. Use cobwebbing to estimate the populations for the next six time increments in each of the following models. (Take the initial populations as indicated by "P_0".)

a.
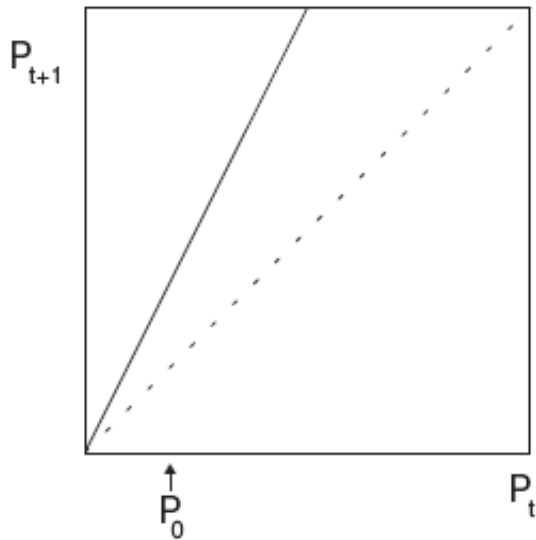
b.
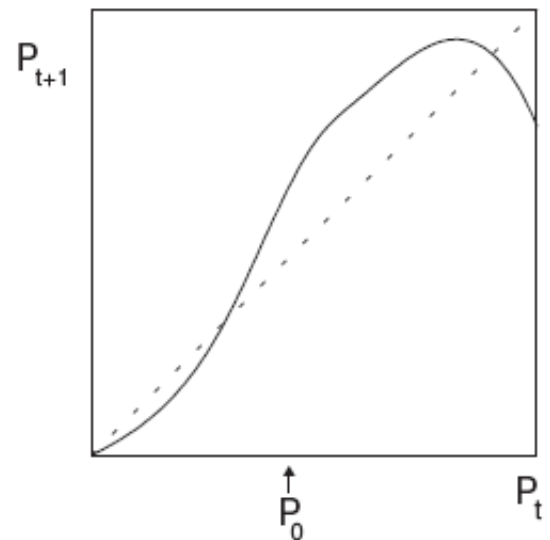
c.
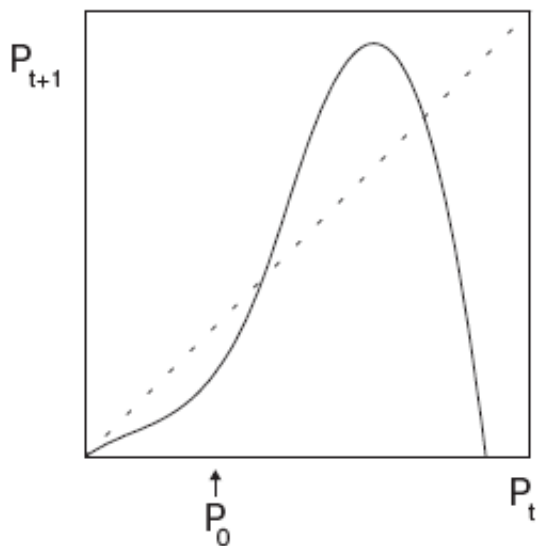
d.
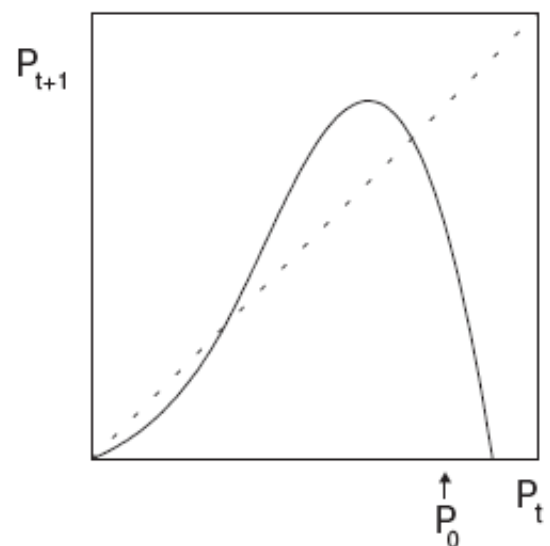
Problems SDE3: Equilibria and linearizations

1. For each of the following, determine the equilibrium points:

   (a) $P^+ = 1.3P - .02P^2$

   (b) $P^+ = 3.2P - .05P^2$

   (c) $\Delta P = .2P(1 - P/20)$

   (d) $\Delta P = \alpha P - \beta P^2$

   (e) $P^+ = \varepsilon P - \delta P^2$.

2. For (a-e) of the preceding problem, linearize the model (i.e,, compute $F'(P)$) first about the steady state 0, and then about the other steady state to determine their respective stabilities. (To compute linearizations, you must convert (c) and (d) to the iteration form "$P^+ = F(P)$".) In parts (d) and (e), your answers will be algebraic conditions on the parameters.

3. Use any software of your choice to

   (a) Plot the first 100 iterates of the logistic iteration with parameter $r = 2.55$ and initial condition (approximately) 0.5;

   (b) You will see that there is a period-8 orbit for this parameter; provide the 8 numbers for this orbit

Problems SDE4: Oscillations, Bifurcations, Chaos

1. The following is a bifurcation diagram for a certain one dimensional iteration that depends on a parameter "r", for r ranging from 1.8 to 2.7. (For your convenience, shown in the right is a zoomed version, for the region where r ranges from 2.4 to 2.6.)
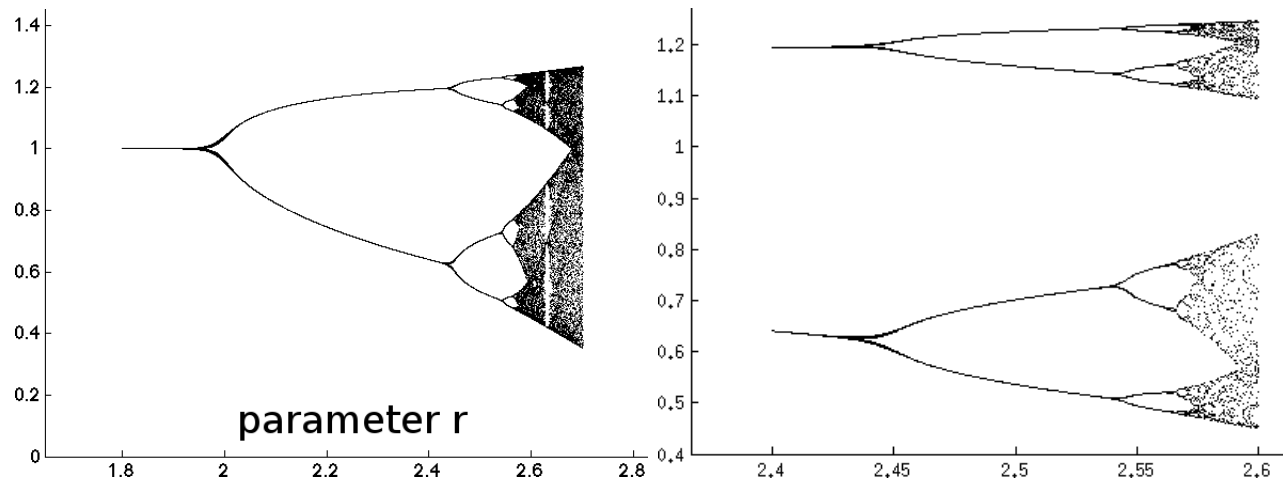
Shown below are iteration plots for the parameters: $r = 1.85, r = 2.2, r = 2.45, r = 2.56, r = 2.7$. Using the information from the bifurcation diagram, you should label each figure that follows by the appropriate value of r. Just write something like "$r = 2.45$" (or "no possible r between 1.8 and 2.7") next to the corresponding figure; no need to provide any justification.
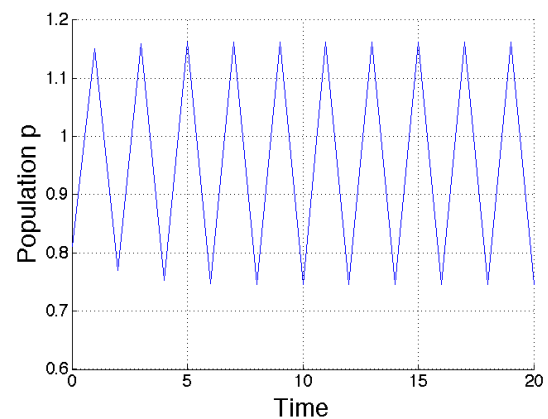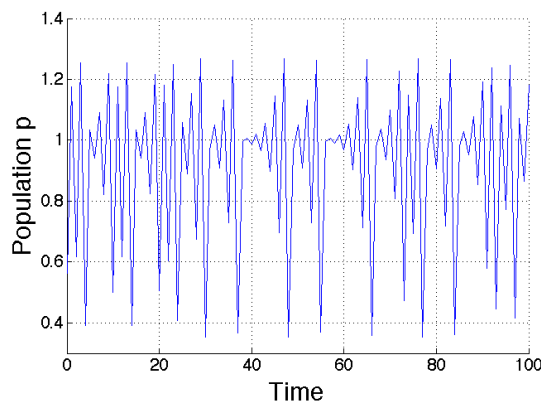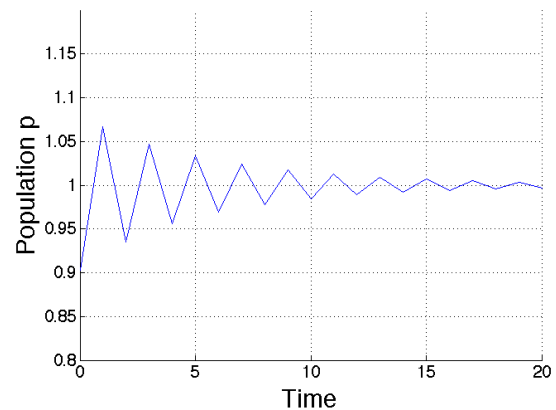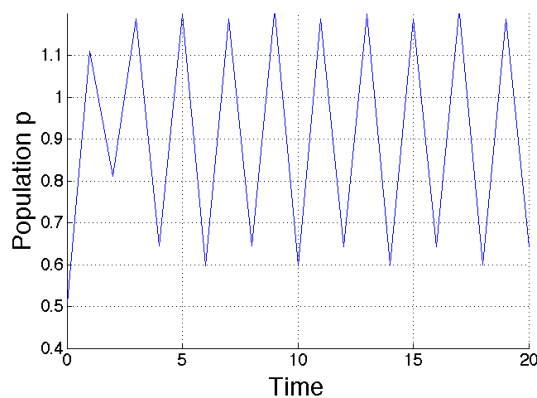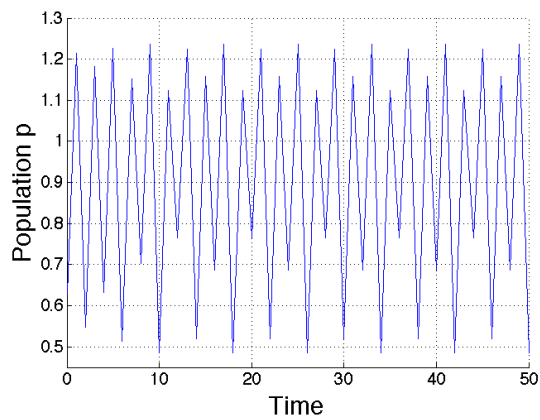
2. Using the same data as in the previous problem, shown below are cobweb plots for the parameters: $r = 1.85, r = 2.2, r = 2.45, r = 2.56, r = 2.7$. Label appropriately.
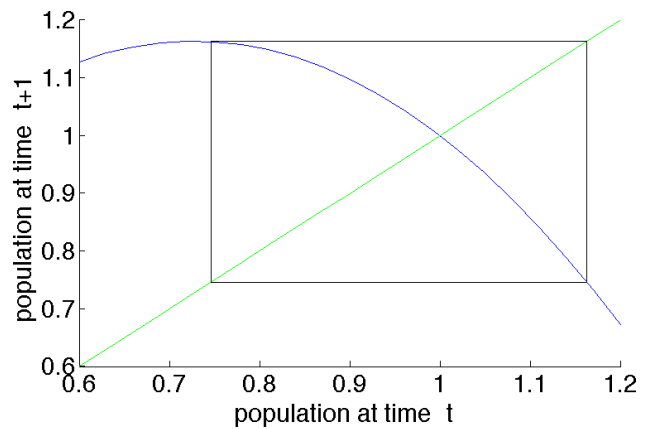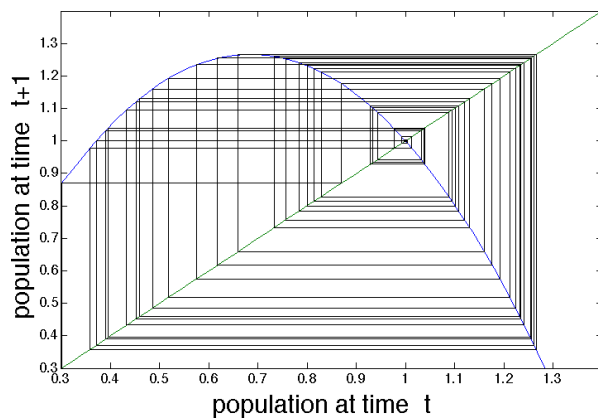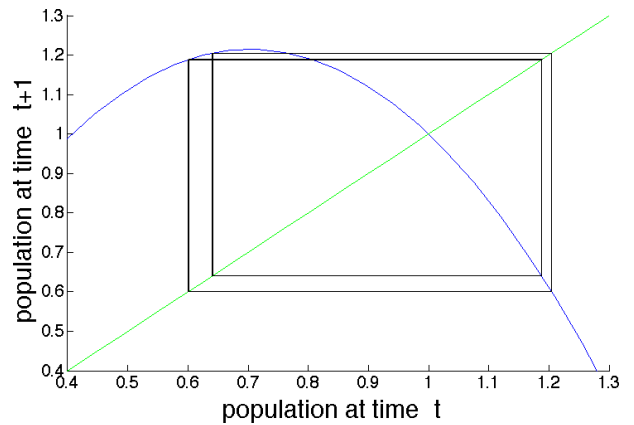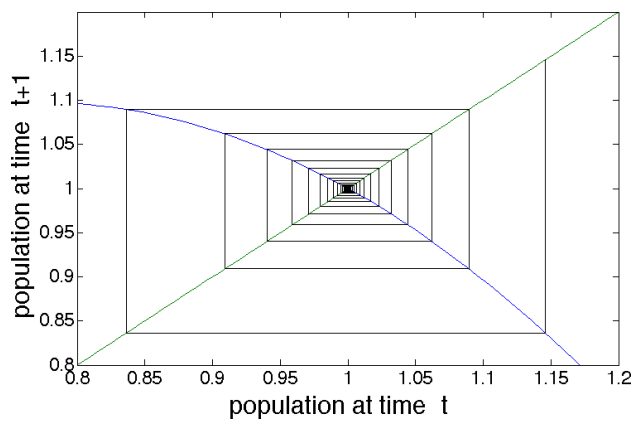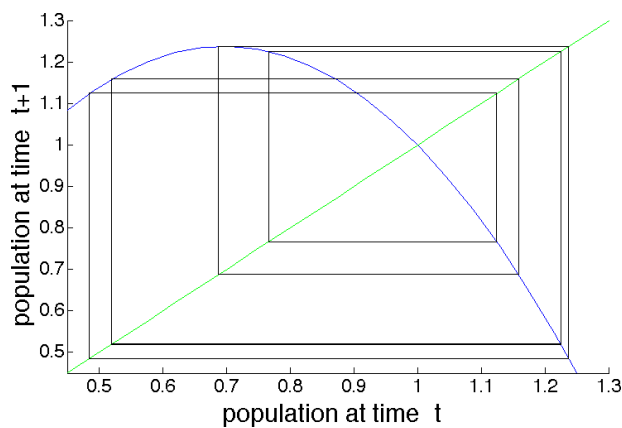
Chapter 2

Deterministic ODE models

2.1 Modeling, Growth, Number of Parameters

Let us start by reviewing a subject treated in basic differential equations courses, namely how one derives differential equations for simple exponential growth and other simple models.

2.1.1 Exponential Growth: Modeling

Suppose that $N(t)$ counts the population of a microorganism in culture, at time t, and write the increment in a time interval $[t, t + h]$ as "$g(N(t), h)$", so that we have:

$$N(t + h) = N(t) + g(N(t), h).$$

(The increment depends on the previous $N(t)$, as well as on the length of the time interval.)

We expand g using a Taylor series to second order:

$$g(N, h) = a + bN + ch + eN^2 + fh^2 + KNh + \text{cubic and higher order terms}$$

($a, b, \ldots$ are some constants). Observe that

$$g(0, h) \equiv 0 \text{ and } g(N, 0) \equiv 0,$$

since there is no increment if there is no population or if no time has elapsed. The first condition tells us that

$$a + ch + fh^2 + \ldots \equiv 0,$$

for all h, so $a = c = f = 0$, and the second condition (check!) says that also $b = N = 0$.
Thus, we conclude that:

$$g(N, h) = KNh + \text{cubic and higher order terms.}$$

So, *for h and N small:*

$$N(t + h) = N(t) + KN(t)h, \tag{2.1}$$

which says that

the increase in population during a (small) time interval
is proportional to the interval length and initial population size.

This means, for example, that if we double the initial population or if we double the interval, the resulting population growth is doubled.

Obviously, (2.1) should not be expected to be true for large h, because of "compounding" effects. It may or may not be true for large N, as we will discuss later.

We next explore the consequences of assuming Equation (2.1) holds for all small $h>0$ and all N.

As usual in applied mathematics, the "*proof is in the pudding*":
one makes such an assumption, explores mathematical consequences that follow from it,
and generates predictions to be *validated experimentally*.
If the predictions pan out, we might want to keep the model.
If they do not, it is back to the drawing board and a new model has to be developed!

2.1.2 Exponential Growth: Math

From our approximation

$$KN(t)h = N(t+h) - N(t)$$

we have that

$$KN(t) = \frac{1}{h}(N(t+h) - N(t))$$

Taking the limit as $h \to 0$, and remembering the definition of derivative, we conclude that the right-hand side converges to $\frac{dN}{dt}(t)$. We conclude that N satisfies the following differential equation:

$$\boxed{\frac{dN}{dt} = KN}.$$

(2.2)

We may solve this equation by the method of *separation of variables*, as follows:

$$\frac{dN}{N} = K\,dt \ \Rightarrow\ \int \frac{dN}{N} = \int K\,dt \ \Rightarrow\ \ln N = Kt + c\,.$$

Evaluating at $t = 0$, we have $\ln N_0 = c$, so that $\ln(N(t)/N_0) = Kt$. Taking exponentials, we have:

$$\boxed{N(t) = N_0 e^{Kt}} \qquad \text{(exponential growth: Malthus, 1798)}$$

Bacterial populations tend to growth exponentially, so long as enough nutrients are available.

2.1.3 Limits to Growth: Modeling

Suppose now there is some number B (the *carrying capacity of the environment*) so that populations $N > B$ are not sustainable, i.e.. $dN/dt < 0$ whenever $N = N(t) > B$:
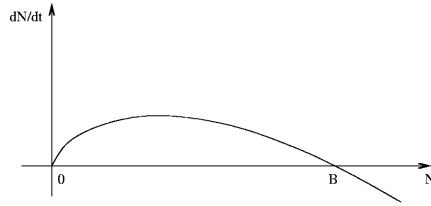
It is reasonable to pick the simplest function that satisfies the stated requirement; in this case, a parabola:

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{B}\right) \qquad \text{(for some constant } r > 0\text{)} \qquad (2.3)$$

But there is a *different way to obtain the same equation*, as follows.
Suppose that the growth rate "K" in Equation (2.2) depends on *availability of a nutrient*:

$$K = K(C) = K(0) + \kappa C + o(C) \approx \kappa C \quad \text{(using that } K(0) = 0\text{)}$$

where $C = C(t)$ denotes the amount of the nutrient, which is depleted in proportion to the population change: [1]

$$\frac{dC}{dt} = -\alpha\frac{dN}{dt} = -\alpha KN$$

("20 new individuals formed $\Rightarrow \alpha \times 20$ less nutrient"). It follows that

$$\frac{d}{dt}(C + \alpha N) = \frac{dC}{dt} + \alpha\frac{dN}{dt} = -\alpha KN + \alpha KN = 0$$

and therefore $C(t) + \alpha N(t)$ must be constant, which we call "C_0"[2]
(we use this notation because $C(0) + \alpha N(0) \approx C(0)$, if the population starts as $N(0) \approx 0$).

So $K = \kappa C = \kappa(C_0 - \alpha N)$, and Equation (2.2) becomes the same equation as (2.3), just with different names of constants:

$$\frac{dN}{dt} = \kappa(C_0 - \alpha N)N$$

2.1.4 Logistic Equation: Math

We solve $\dfrac{dN}{dt} = rN\left(1 - \dfrac{N}{B}\right) = r\dfrac{N(B-N)}{B}$ using again the method of separation of variables:

$$\int \frac{B\,dN}{N(B-N)} = \int r\,dt\,.$$

We compute the integral using a partial fractions expansion:

$$\int\left(\frac{1}{N} + \frac{1}{B-N}\right)dN = \int r\,dt \;\Rightarrow\; \ln\left(\frac{N}{B-N}\right) = rt+c \;\Rightarrow\; \frac{N}{B-N} = \widetilde{c}e^{rt} \;\Rightarrow\; N(t) = \frac{\widetilde{c}B}{\widetilde{c} + e^{-rt}}$$

---

[1] if $N(t)$ counts the number of individuals, this is somewhat unrealistic, as it the ignores depletion of nutrient due to the growth or individuals once they are born; it is sometimes better to think of $N(t)$ as the *total biomass* at time t
[2] this is an example of a "conservation law", as we'll discuss later

$$\Rightarrow \; \widetilde{c} = N_0/(B - N_0) \quad \Rightarrow \quad \boxed{N(t) \; = \; \frac{N_0 B}{N_0 + (B - N_0)e^{-rt}}}$$

We can see that there is a B asymptote as $t \to \infty$. Let's graph with Maple:

```
with(plots):
f(t):=t->(0.2)/(0.2+0.8*exp(-t)):
p1:=plot(f(t),0..8,0..1.3,tickmarks=[0,2],thickness=3,color=black):
g:=t->1:
p2:=plot(g(t),0..8,tickmarks=[0,2],thickness=2,linestyle=2,color=black):
display(p1,p2);
```
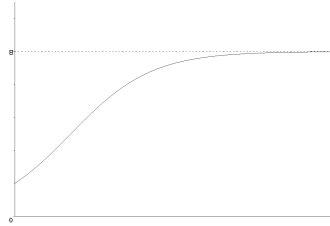

Gause's 1934 Experiments

G.F. Gause carried out experiments in 1934, involving Paramecium caudatum and Paramecium aurelia, which show clearly logistic growth:
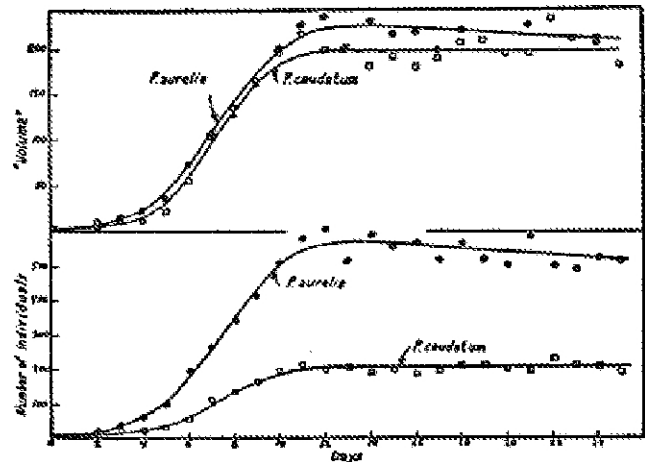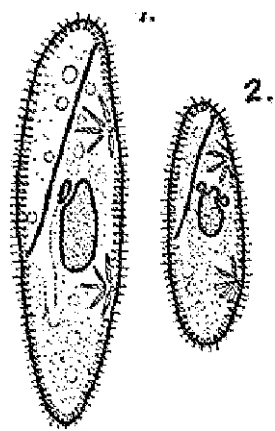
(# individuals and volume of P. caudatum and P. aurelia, cultivated separately, medium changed daily, 25 days.)

2.1.5 Changing Variables, Rescaling Time

We had this equation for growth under nutrient limitations:

$$\boxed{\frac{dN}{dt} \; = \; \kappa \left(C_0 - \alpha N\right) N}$$

which we solved explicitly (and graphed for some special values of the parameters C_0, κ, α).
But how do we know that "qualitatively" the solution "looks the same" for other parameter values?

Can the *qualitative* behavior of solutions depend upon the actual numbers C_0, κ, α?

First of all, we notice that we could collect terms as

$$\frac{dN}{dt} = ((\kappa C_0) - (\kappa \alpha)N) N = \left(\widetilde{C}_0 - \widetilde{\alpha} N\right) N$$

(where $\widetilde{C}_0 = \kappa C_0$ and $\widetilde{\alpha} = \kappa \alpha$), so that we might as well suppose that $\kappa = 1$ (but change α, C_0).

But we can do even better and use changes of variables in N and t in order to eliminate the two remaining parameters!

The idea is as follows. Suppose that $N(t)$ is any solution of the differential equation

$$\frac{dN}{dt} = f(t, N(t))$$

(we allow an explicit dependence of f on t in order to make the explanation more general, even though most examples given below do not show an explicit t). Let us now introduce a new function, called N^*, that depends on a new time variable, called t^*, by means of the following definition:

$$N^*(t^*) := \frac{1}{\hat{N}} N(t^* \hat{t})$$

where $\hat{N}$ and $\hat{t}$ are two constants. These two constants will be specified later; we will pick them in such a way that the equations will end up having fewer parameters. The chain rule says that:

$$\frac{dN^*}{dt^*}(t^*) = \frac{\hat{t}}{\hat{N}} \frac{dN}{dt}(t^* \hat{t}) = \frac{\hat{t}}{\hat{N}} f(t^* \hat{t}, N(t^* \hat{t})).$$

(The expression "dN/dt" above might be confusing, but it should not be. We are simply writing "$dN/dt(t^* \hat{t})$" instead of "$N'(t^* \hat{t})$". The "t" variable is a dummy variable in this expression.) In summary, we may symbolically write:

$$`` \frac{dN}{dt} = \frac{d(N^* \hat{N})}{d(t^* \hat{t})} = \frac{\hat{N}}{\hat{t}} \frac{dN^*}{dt^*} \text{ ,''}$$

and proceed formally. Our general strategy will be:

- Write each variable (in this example, N and t) as a product of a new variable and a still-to-be-determined constant.
- Substitute into the equations, simplify, and collect terms.
- Finally, pick values for the constants so that the equations (in this example, there is only one differential equation, but in other examples there may be several) have as few remaining parameters as possible.

The procedure can be done in many ways (depending on how you collect terms, etc.), so different people may get different solutions.

Let's follow the above procedure with our example. We start by writing: $N = N^* \hat{N}$ and $t = t^* \hat{t}$, *where stars indicate new variables and the hats are constants to be chosen.* Proceeding purely formally, we substitute these into the differential equation:

$$\frac{d\left(N^* \hat{N}\right)}{d\left(t^* \hat{t}\right)} = \kappa \left(C_0 - \alpha N^* \hat{N}\right) N^* \hat{N} \quad \Rightarrow \quad \frac{\hat{N}}{\hat{t}} \frac{dN^*}{dt^*} = \kappa \left(C_0 - \alpha N^* \hat{N}\right) N^* \hat{N}$$

$$\Rightarrow \quad \frac{dN^*}{dt^*} = \kappa \hat{t} \alpha \hat{N} \left(\frac{C_0}{\alpha \hat{N}} - N^*\right) N^*$$

Let us look at this last equation: we'd like to make $\frac{C_0}{\alpha \hat{N}} = 1$ and $\kappa \hat{t} \alpha \hat{N} = 1$.

But this can be done! Just pick: $\boxed{\hat{N} := \dfrac{C_0}{\alpha}}$ and $\hat{t} = \dfrac{1}{\kappa \alpha \hat{N}}$, that is: $\boxed{\hat{t} := \dfrac{1}{\kappa C_0}}$

$$\rightsquigarrow \quad \frac{dN^*}{dt} = (1 - N^*)\, N^* \quad \text{or, drop stars, and write just} \quad \frac{dN}{dt} = (1 - N)\, N$$

Thus, we can analyze the simpler system.

However, we should remember that the new "N" and "t" are rescaled versions of the old ones. In order to understand how to bring everything back to the original coordinates, note that another way to express the relation between N and N^* is as follows:

$$N(t) \;=\; \hat{N} N^* \left(\frac{t}{\hat{t}} \right)$$

This formula allows us to recover the solution $N(t)$ to the original problem once that we have obtained the solution to the problem in the N^*, t^* coordinates. Concretely, in our example, as $t/\hat{t} = t/\frac{1}{\kappa C_0}$:

$$N(t) \;=\; \frac{C_0}{\alpha} N^* \left(\kappa C_0 t \right)$$

If we have a plotted solution of the equation $\frac{dN^*}{dt^*} = (1 - N^*)\, N^*$ then the plot in original variables is obtained by stretching or contracting the plot in these new variables.

We may think of N^*, t^* as quantity & time in some new units of measurement. This procedure is related to "nondimensionalization" of equations, which we'll mention later.

2.1.6 A More Interesting Example: the Chemostat
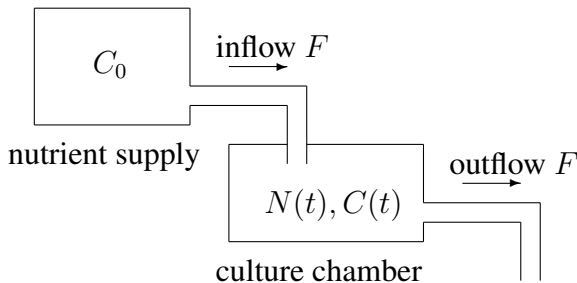
$V = $ constant volume of solution in culture chamber
$F = $ (constant and equal) flows in vol/sec, e.g. m^3/s
$N(t) = $ bacterial concentration in mass/vol, e.g. g/m^3
$C_0, C(t) = $ nutrient concentrations in mass/vol
    (C_0 assumed constant)
chamber is well-mixed
("continuously stirred tank reactor (CSTR)" in chem engr)

Assumptions (same as in second derivation of logistic growth):

- growth of biomass in each unit of volume proportional to population (and to interval length), and depends on amount of nutrient in that volume (we think of density as the mass in a small unit volume):

$$N(t + \Delta t) - N(t) \quad \text{due to growth} \quad = \quad K(C(t))\, N(t)\, \Delta t$$

(the choice of the function $K(C)$ is discussed below)

- consumption of nutrient per unit volume proportional to increase of bacterial population:

$$C(t + \Delta t) - C(t) \quad \text{due to consumption} \quad = \quad -\alpha \left[N(t + \Delta t) - N(t) \right]$$

2.1.7 Chemostat: Mathematical Model

total biomass: $N(t)\,V$ and   total nutrient in culture chamber: $C(t)\,V$

biomass change in interval Δt due to growth:

$$N(t+\Delta t)V - N(t)V \;=\; [N(t+\Delta t) - N(t)]V \;=\; K(C(t))\,N(t)\,\Delta t\,V$$

so contribution to $d(NV)/dt$ is "$+K(C)NV$"

bacterial mass in effluent:
in a small interval Δt, the volume out is: $F \cdot \Delta t\ (\frac{m^3}{s}s =)m^3$
so, since the concentration is $N(t)\ g/m^3$, the mass out is: $N(t) \cdot F \cdot \Delta t\ g$
and so the contribution to $d(NV)/dt$ is "$-N(t)F$"

for $d(CV)/dt$ equation:
we have three terms: $-\alpha K(C)NV$ (depletion), $-C(t)F$ (outflow), and $+C_0 F$ (inflow), $\rightsquigarrow$

$$\frac{d(NV)}{dt} \;=\; K(C)NV - NF$$
$$\frac{d(CV)}{dt} \;=\; -\alpha K(C)NV - CF + C_0 F\,.$$

Finally, divide by the constant V to get this system of equations on N, C:

$$\frac{dN}{dt} \;=\; K(C)N - NF/V$$
$$\frac{dC}{dt} \;=\; -\alpha K(C)N - CF/V + C_0 F/V$$

2.1.8 Michaelis-Menten Kinetics

A reasonable choice for "$K(C)$" is as follows (later, we come back to this topic in much more detail):

$$K(C) \;=\; \frac{k_{\max}\, C}{k_n + C} \quad \text{or, in another very usual notation:} \quad \frac{V_{\max}\, C}{K_{\mathrm{m}} + C}\,.$$
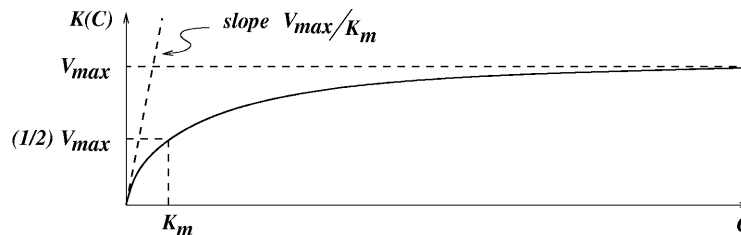
This gives linear growth for small nutrient concentrations:

$$K(C) \approx K(0) + K'(0)C \;=\; \frac{V_{\max}\, C}{K_{\mathrm{m}}}$$

but saturates at $V_{\max}$ as $C \to \infty$.

(More nutrient $\Rightarrow$ more growth, but only up to certain limits — think of a buffet dinner!)

Note that when $C = K_{\mathrm{m}}$, the growth rate is 1/2 ("m" for middle) of maximal, i.e. $V_{\mathrm{max}}/2$,

We thus have these equations for the chemostat with MM Kinetics:

$$\frac{dN}{dt} = \frac{k_{\mathrm{max}} C}{k_n + C} N - (F/V)N$$

$$\frac{dC}{dt} = -\alpha \frac{k_{\mathrm{max}} C}{k_n + C} N - (F/V)C + (F/V)C_0$$

Our next goal is to study the behavior of this system of two ODE's
for all possible values of the six parameters $k_{\mathrm{max}}, k_n, F, V, C_0, \alpha$.

2.1.9   Side Remark: "Lineweaver-Burk plot" to Estimate Parameters

Suppose we measured experimentally $K(C_i)$ for various values C_i.
How does one estimate K_{m} and V_{max}?

Solution: observe that

$$\frac{1}{K(C)} = \frac{K_{\mathrm{m}} + C}{V_{\mathrm{max}} C} = \frac{1}{V_{\mathrm{max}}} + \frac{K_{\mathrm{m}}}{V_{\mathrm{max}}} \cdot \frac{1}{C}$$

therefore, $1/K(C)$ is a *linear* function of $1/C$!
Thus, just plot $1/K(C)$ against $1/C$ and fit a line (linear regression).
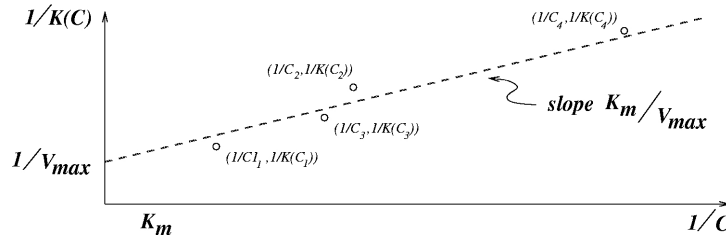
2.1.10   Chemostat: Reducing Number of Parameters

Following the procedure outlined earlier, we write: $C = C^* \hat{C}, \; N = N^* \hat{N}, \; t = t^* \hat{t}$, and substitute:

$$\frac{d(N^* \hat{N})}{d(t^* \hat{t})} = \frac{k_{\mathrm{max}} C^* \hat{C}}{k_n + C^* \hat{C}} N^* \hat{N} - (F/V)N^* \hat{N}$$

$$\frac{d(C^* \hat{C})}{d(t^* \hat{t})} = -\alpha \frac{k_{\mathrm{max}} C^* \hat{C}}{k_n + C^* \hat{C}} N^* \hat{N} - (F/V)C^* \hat{C} + (F/V)C_0$$

$$\frac{dN}{dt} = \frac{d(N^* \hat{N})}{d(t^* \hat{t})} = \frac{\hat{N}}{\hat{t}} \frac{dN^*}{dt^*} \; \& \; \frac{dC}{dt} = \frac{d(C^* \hat{C})}{d(t^* \hat{t})} = \frac{\hat{C}}{\hat{t}} \frac{dC^*}{dt^*} \rightsquigarrow$$

$$\frac{dN^*}{dt^*} = \frac{\hat{t} \, k_{\mathrm{max}} C^* \hat{C}}{k_n + C^* \hat{C}} N^* - \frac{\hat{t} F}{V} N^*$$

$$\frac{dC^*}{dt^*} = -\alpha \frac{\hat{t} \, k_{\mathrm{max}} C^*}{k_n + C^* \hat{C}} N^* \hat{N} - \frac{\hat{t} F}{V} C^* + \frac{\hat{t} F}{\hat{C} V} C_0$$

or equivalently:

$$\frac{dN^*}{dt^*} = (\hat{t}\, k_{\max})\frac{C^*}{k_n/\hat{C} + C^*}\, N^* - \frac{\hat{t}F}{V}N^*$$

$$\frac{dC^*}{dt^*} = -\left(\frac{\alpha\hat{t}\, k_{\max}\hat{N}}{\hat{C}}\right)\frac{C^*}{k_n/\hat{C} + C^*}\, N^* - \frac{\hat{t}F}{V}C^* + \frac{\hat{t}F}{\hat{C}V}C_0$$

It would be nice, for example, to make $k_n/\hat{C} = 1$, $\dfrac{\hat{t}F}{V} = 1$, and $\dfrac{\alpha\hat{t}\, k_{\max}\hat{N}}{\hat{C}} = 1$. This can indeed be done, provided that we define: $\hat{C} := k_n$, $\quad \hat{t} := \dfrac{V}{F}$, $\quad$ and $\hat{N} := \dfrac{\hat{C}}{\alpha\hat{t}\, k_{\max}} = \dfrac{k_n}{\alpha\hat{t}\, k_{\max}} = \dfrac{k_n F}{\alpha V\, k_{\max}}$

$$\rightsquigarrow \quad \frac{dN^*}{dt^*} = \left(\frac{V k_{\max}}{F}\right)\frac{C^*}{1 + C^*}\, N^* - N^*$$

$$\frac{dC^*}{dt^*} = -\frac{C^*}{1 + C^*}\, N^* - C^* + \frac{C_0}{k_n}$$

or, introducing two new constants $\alpha_1 = \left(\frac{V k_{\max}}{F}\right)$ and $\alpha_2 = \dfrac{C_0}{k_n}$ we end up with:

$$\frac{dN^*}{dt^*} = \alpha_1 \frac{C^*}{1 + C^*}\, N^* - N^*$$

$$\frac{dC^*}{dt^*} = -\frac{C^*}{1 + C_\star}\, N^* - C^* + \alpha_2$$

We will study how the behavior of the chemostat depends on these two parameters, always remembering to "translate back" into the original parameters and units.

The old and new variables are related as follows:

$$N(t) = \hat{N}N^*\left(\frac{t}{\hat{t}}\right) = \frac{k_n F}{\alpha V\, k_{\max}}N^*\left(\frac{F}{V}t\right), \qquad C(t) = \hat{C}C^*\left(\frac{t}{\hat{t}}\right) = k_n C^*\left(\frac{F}{V}t\right)$$

Remark on units

Since $k_{\max}$ is a rate (obtained at saturation), it has units time^{-1}; thus, α_1 is "dimensionless". Similarly, k_n has units of concentration (since it is being added to C, and in fact for $C = k_n$ we obtain half of the max rate $k_{\max}$), so also α_2 is dimensionless.

Dimensionless constants are a nice thing to have, since then we can talk about their being "small" or "large". (What does it mean to say that a person of height 2 is tall? 2 cm? 2in? 2 feet? 2 meters?) We do not have time to cover the topic of units and non-dimensionalization in this course, however.

2.1.11   Signs of interactions among variables

It is often convenient to classify systems according to the signs of the interactions between pairs of variables. To keep matters simple, let us restrict ourselves here to systems consisting of two components x and y. These components might represent the concentrations of certain intracellular chemicals –for instance, of two proteins– in molecular biology, or, in the context of ecology, the numbers of individuals of a certain pair of species.

So we have a two-dimensional system of ODE's with variables $x(t)$ and $y(t)$:

$$\frac{dx}{dt} = f(x,y)$$
$$\frac{dy}{dt} = g(x,y) \,.$$

We will suppose that the signs of the partial derivatives

$$\frac{\partial f}{\partial y}(x,y) \quad \text{and} \quad \frac{\partial g}{\partial x}(x,y)$$

are either always ≥ 0 or always ≤ 0 when evaluated at all values (x, y) of interest (typically, $x \geq 0$ and $y \geq 0$).[3] It is useful to associate a graph to a system as follows.

Introduce two nodes, called X and Y, and:

1. draw a positive arrow from X to Y "$X \to Y$" provided that $\frac{\partial g}{\partial x}(x,y) > 0$,

2. draw a blunt arrow from X to Y "$X \dashv Y$" provided that $\frac{\partial g}{\partial x}(x,y) < 0$,

3. draw a positive arrow from Y to X "$Y \to X$" provided that $\frac{\partial f}{\partial y}(x,y) > 0$,

4. draw a blunt arrow from Y to X "$Y \dashv X$" provided that $\frac{\partial f}{\partial y}(x,y) < 0$

(no arrow is drawn if a partial derivative is identically zero). When analyzing examples, there is often no need to take derivatives. For instance, suppose that

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

(obviously a contrived mathematical example!). It is clear that an increasing y makes f decrease (because the denominator in the exponent is an increasing function of y, and the exponential is an increasing function), so $Y \dashv X$.

Assuming that both partial derivatives are everywhere nonzero (and have constant sign), there are three types of interactions, graphically shown here:

---

[3]We do not analyze what happens if a partial derivative is sometimes positive and sometimes negative. For example, if $g(x,y) = (x - x^2)y$, then $dy/dt > 0$ when $0 < x < 1$, but $dy/dt > 0$ when $x > 1$. For example, x might be a nutrient that is needed for the growth of y, but which is toxic when overdosing.

Mutualism
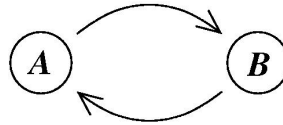
This happens if the variables have positive ("activation") effects on each other's growth rates. A positive change in A results in a positive change in the growth of B, and vice-versa. These interactions create a positive feedback on both variables.

Configurations like these are associated to signal amplification and production of switch-like biochemical responses.

Competition or mutual inhibition
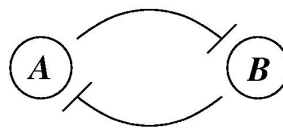
This happens if the variables have negative ("repression") effects on each other's growth rates. A positive change in A results in repression of growth of B, and repression of B in turn enhances the growth of A. These interactions also create a positive feedback on both variables.

Such configurations allow systems to exhibit multiple discrete, alternative stable steady-states, thus providing a mechanism for memory. They also help in allowing sharp ("binary") responses to inputs and are important in cell decision-making (apoptosis, division, . . .).

Activation-inhibition or predator-prey
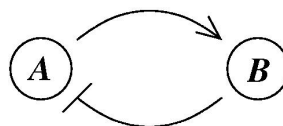
This happens if the variables have opposite effects on each other's growth rates. A negative feedback is created. Activation-inhibition configurations of this type are necessary for the generation of periodic behaviors such as circadian rhythms or cell cycle oscillations, as well as for tight regulation (homeostasis) of physiological variables.

Monotone systems

The first two types of systems, mutualistic (or, as is said in mathematical biology, "cooperative"), and mutually inhibiting ("competitive") are examples of what are called "monotone systems," for which a rich theory (and not just for 2-d systems) exists[4]. Let us illustrate a simple result for the mutually activating case. (A similar result holds true for mutual repression.)

---

[4]For instance, see an intuitive discussion of results in: E.D. Sontag. Monotone and near-monotone biochemical networks. Systems and Synthetic Biology, 1:59-87, 2007.
http://www.springerlink.com/content/x44774083876j062/fulltext.pdf

Claim: there cannot be any periodic orbit in such a system.

Proof by contradiction: Suppose that there would be a periodic orbit in which the motion is counter-clockwise, as shown in the left part of this figure:
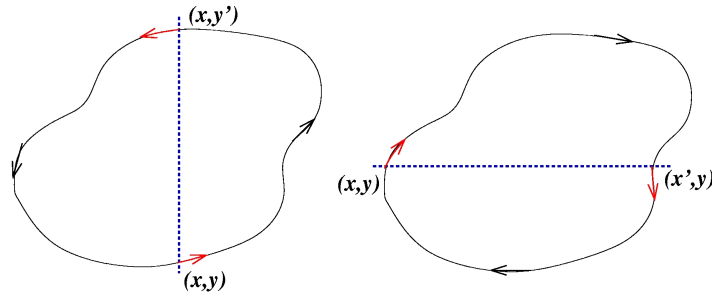
We now pick two points in this orbit with identical x coordinates, as indicated by (x, y) and (x, y') in the figure. These points correspond to the concentrations at two times t_0, t_1, with $x(t_0) = x(t_1)$ and $y(t_0) < y(t_1)$. Since $y(t_1)$ is larger than $y(t_0)$, x is at the same concentration, and the species are mutually activating, it follows that the rate of change in the concentration x should be comparatively larger at time t_1 than at time t_0, that is, $f(x, y') \geq f(x, y)$. However, this contradicts the fact that $x(t)$ is increasing at time t_0 ($f(x, y) \geq 0$) but is decreasing at time t_1 ($f(x, y') \leq 0$). The contradiction means that there cannot be any counterclockwise-oriented curve. To show that there cannot be any clockwise-oriented curve, one may proceed by an entirely analogous argument, using two points (x, y) and (x', y) as in the figure.

2.2 Steady States and Linearized Stability Analysis

2.2.1 Steady States

The key to the "geometric" analysis of systems of ODE's is to write them in vector form:

$$\frac{dX}{dt} = F(X) \qquad \text{(where } F \text{ is a vector function and } X \text{ is a vector)}.$$

The vector $X = X(t)$ has some number n of components, each of which is a function of time. One writes the components as x_i ($i = 1, 2, 3, \ldots, n$), or when $n = 2$ or $n = 3$ as x, y or x, y, z, or one uses notations that are related to the problem being studied, like N and C for the concentration (or the biomass) of a population and C for the concentration of a nutrient.
For example, the chemostat

$$\frac{dN}{dt} = \alpha_1 \frac{C}{1+C} N - N$$
$$\frac{dC}{dt} = -\frac{C}{1+C} N - C + \alpha_2$$

may be written as $\dfrac{dX}{dt} = F(X) = \begin{pmatrix} f(N,C) \\ g(N,C) \end{pmatrix}$, provided that we define:

$$f(N,C) = \alpha_1 \frac{C}{1+C} N - N$$
$$g(N,C) = -\frac{C}{1+C} N - C + \alpha_2.$$

By definition, a *steady state* or *equilibrium*[5] is any root of the algebraic equation

$$F(\bar{X}) = 0$$

that results when we set the right-hand side to zero.

For example, for the chemostat, a steady state is the same thing as a solution $X = (N, C)$ of the two simultaneous equations

$$\alpha_1 \frac{C}{1+C} N - N = 0$$
$$-\frac{C}{1+C} N - C + \alpha_2 = 0.$$

Let us find the equilibria for this example.

A trick which sometimes works for chemical and population problems, is as follows.
We factor the first equation:

$$\left(\alpha_1 \frac{C}{1+C} - 1 \right) N = 0.$$

---

[5]the word "equilibrium" is used in mathematics as a synonym for steady state, but the term has a more restrictive meaning for physicists and chemists

So, for an equilibrium $\bar{X} = (\bar{N}, \bar{C})$,

$$\text{either} \quad \bar{N} = 0 \quad \text{or} \quad \alpha_1 \frac{\bar{C}}{1 + \bar{C}} = 1 \, .$$

We consider each of these two possibilities separately.

In the first case, $\bar{N} = 0$. Since also it must hold that

$$-\frac{\bar{C}}{1 + \bar{C}} \bar{N} - \bar{C} + \alpha_2 = -\bar{C} + \alpha_2 = 0 \, ,$$

we conclude that $\bar{X} = (0, \alpha_2)$ (no bacteria alive, and nutrient concentration α_2).

In the second case, $\bar{C} = \frac{1}{\alpha_1 - 1}$, and therefore the second equation gives $\bar{N} = \alpha_1 \left(\alpha_2 - \frac{1}{\alpha_1 - 1} \right)$ (check!).

So we found two equilibria:

$$\bar{X}_1 = (0, \alpha_2) \quad \text{and} \quad \bar{X}_2 = \left(\alpha_1 \left(\alpha_2 - \frac{1}{\alpha_1 - 1} \right), \frac{1}{\alpha_1 - 1} \right) \, .$$

However, observe that an equilibrium is physically meaningful only if $\bar{C} \geq 0$ and $\bar{N} \geq 0$. *Negative populations or concentrations, while mathematically valid, do not represent physical solutions.*[6]
The first steady state is always well-defined in this sense, but not the second.
This equilibrium $\bar{X}_2$ is well-defined and makes physical sense only if

$$\alpha_1 > 1 \quad \text{and} \quad \alpha_2 > \frac{1}{\alpha_1 - 1} \tag{2.4}$$

or equivalently:

$$\alpha_1 > 1 \quad \text{and} \quad \alpha_2(\alpha_1 - 1) > 1 \, . \tag{2.5}$$

Reducing the number of parameters to just two (α_1 and α_2) allowed us to obtain this very elegant and compact condition. *But this is not a satisfactory way to explain our conclusions, because α_1, α_2 were only introduced for mathematical convenience, but were not part of the original problem.*

Since, $\hat{t} := \frac{V}{F}$, $\alpha_1 = \hat{t} \, k_{\max} = \frac{V}{F} k_{\max}$ and $\alpha_2 = \frac{\hat{t}F}{\hat{C}V} C_0 = \frac{C_0}{\hat{C}} = \frac{C_0}{k_n}$, the conditions are:

$$k_{\max} > \frac{F}{V} \quad \text{and} \quad C_0 > \frac{k_n}{\frac{V}{F} k_{\max} - 1} \, .$$

The first condition means roughly that the maximal possible bacterial reproductive rate is larger than the tank emptying rate, which makes intuitive sense. *As an exercise, you should similarly interpret "in words" the various things that the second condition is saying.*

Meaning of Equilibria: If a point $\bar{X}$ is an equilibrium, then the constant vector $X(t) \equiv \bar{X}$ is a solution of the system of ODE's, because a constant has zero derivative: $d\bar{X}/dt = 0$, and since $F(\bar{X}) = 0$ by definition of equilibrium, we have that $d\bar{X}/dt = F(\bar{X})$.
Conversely, if a constant vector $X(t) \equiv \bar{X}$ is a solution of $dX(t)/dt = F(X(t))$, then, since $(d/dt)(X(t)) \equiv 0$, also then $F(\bar{X}) = 0$ and therefore $\bar{X}$ is an equilibrium.
In other words, an equilibrium is a point where the solution stays forever.
As you studied in your ODE class, an equilibrium may be stable or unstable (think of a pencil perfectly balanced on the upright position). We next review stability.

---

[6]Analogy: we are told that the length L of some object is a root of the equation $L^2 - 4 = 0$. We can then conclude that the length must be $L = 2$, since the other root, $L = -2$, cannot correspond to a length.

2.2.2   Linearization

We wish to analyze the behavior of solutions of the ODE system $dX/dt = F(X)$ *near a given steady state* $\bar{X}$. For this purpose, it is convenient to introduce the displacement (translation) relative to $\bar{X}$:

$$\hat{X} = X - \bar{X}$$

and to write an equation for the variables $\hat{X}$. We have:

$$\frac{d\hat{X}}{dt} \;=\; \frac{dX}{dt} - \frac{d\bar{X}}{dt} \;=\; \frac{dX}{dt} - 0 \;=\; \frac{dX}{dt} \;=\; F(\hat{X} + \bar{X}) \;=\; \underbrace{F(\bar{X})}_{=0} + F'(\bar{X})\hat{X} + \underbrace{o(\hat{X})}_{\approx 0} \;\approx\; A\hat{X}$$

where $A = F'(\bar{X})$ is the *Jacobian* of F evaluated at $\bar{X}$.
We dropped higher-order-than-linear terms in $\hat{X}$ because we are only interested in $\hat{X} \approx 0$
(small displacements $X \approx \bar{X}$ from $\bar{X}$ are the same as small $\hat{X}$'s).

Recall that the Jacobian, or "derivative of a vector function," is defined as the $n \times n$ matrix whose (i, j)th entry is $\partial f_i / \partial x_j$, if f_i is the ith coordinate of F and x_j is the jth coordinate of x.

One often drops the "hats" and writes the above *linearization* simply as $dX/dt = AX$,
but it is extremely important to remember that what this equation represents:
it is an equation for the displacement from a particular equilibrium $\bar{X}$.
More precisely, it is an equation for *small* displacements from $\bar{X}$.
(And, for any other equilibrium $\bar{X}$, a different matrix A will, generally speaking, result).

For example, let us take the chemostat, after a reduction of the number of parameters:

$$\frac{d}{dt}\begin{pmatrix} N \\ C \end{pmatrix} \;=\; F(N, C) \;=\; \begin{pmatrix} \alpha_1 \frac{C}{1+C} N - N \\ -\frac{C}{1+C} N - C + \alpha_2 \end{pmatrix}$$

so that, at any point (N, C) the Jacobian $A = F'$ of F is:

$$\begin{pmatrix} \alpha_1 \frac{C}{1+C} - 1 & \frac{\alpha_1 N}{(1+C)^2} \\ -\frac{C}{1+C} & -\frac{N}{(1+C)^2} - 1 \end{pmatrix}.$$

In particular, at the point $\bar{X}_2$, where $\bar{C} = \frac{1}{\alpha_1 - 1}$, $\bar{N} = \frac{\alpha_1(\alpha_1\alpha_2 - \alpha_2 - 1)}{\alpha_1 - 1}$ we have:

$$\begin{bmatrix} 0 & \beta(\alpha_1 - 1) \\ -\dfrac{1}{\alpha_1} & -\dfrac{\beta(\alpha_1 - 1) + \alpha_1}{\alpha_1} \end{bmatrix}$$

where we used the shorthand: $\beta = \alpha_2(\alpha_1 - 1) - 1$. (Prove this as an exercise!)

Remark. An important result, the Hartman-Grobman Theorem, justifies the study of linearizations. It states that solutions of the nonlinear system $\frac{dX}{dt} = F(X)$ in the vicinity of the steady state $\bar{X}$ look "qualitatively" just like solutions of the linearized equation $dX/dt = AX$ do in the vicinity of the point $X = 0$.[7]

For linear systems, stability may be analyzed by looking at the eigenvalues of A, as we see next.

---

[7]The theorem assumes that none of the eigenvalues of A have zero real part ("hyperbolic fixed point"). "Looking like" is defined in a mathematically precise way using the notion of "homeomorphism" which means that the trajectories look the same after a continuous invertible transformation, that is, a sort of "nonlinear distortion" of the phase space.

2.2.3   Review of (Local) Stability

For the purposes of this course, we'll say that a linear system $dX/dt = AX$, where A is $n \times n$ matrix, is *stable* if all solutions $X(t)$ have the property that $X(t) \to 0$ as $t \to \infty$. The main theorem is:

> *stability is equivalent to: the real parts of all the eigenvalues of A are negative*

For nonlinear systems $dX/dt = F(X)$, one applies this condition as follows:[8]

- For each steady state $\bar{X}$, compute A, the Jacobian of F evaluated at $\bar{X}$, and test its eigenvalues.

- If *all* the eigenvalues of A have negative real part, conclude *local stability*:
  every solution of $dX/dt = F(X)$ that starts near $X = \bar{X}$ converges to $\bar{X}$ as $t \to \infty$.

- If A has even one eigenvalue with positive real part, then the corresponding nonlinear system $dX/dt = F(X)$ is *unstable* around $\bar{X}$, meaning that at least some solutions that start near $\bar{X}$ will move away from $\bar{X}$.

The linearization $dX/dt = AX$ at a steady state $\bar{X}$ *says nothing at all about global stability, that is to say, about behaviors of $dX/dt = F(X)$ that start at initial conditions that are far away from $\bar{X}$.*
For example, compare the two equations: $dx/dt = -x - x^3$ and $dx/dt = -x + x^2$.
In both cases, the linearization at $x = 0$ is just $dx/dt = -x$, which is stable.
In the first case, it turns out that all the solutions of the nonlinear system also converge to zero.
(Just look at the phase line.)
However, in the second case, even though the linearization is the same, it is not true that all solutions converge to zero. For example, starting at a state $x(0) > 1$, solutions diverge to $+\infty$ as $t \to \infty$.
(Again, this is clear from looking at the phase line.)

It is often confusing to students that from the fact that all solutions of $dX/dt = AX$ converge to *zero*, one concludes for the nonlinear system that all solutions converge to $\bar{X}$.
The confusion is due simply to notations: we are really studying $d\hat{X}/dt = A\hat{X}$, where $\hat{X} = X - \bar{X}$, but we usually drop the hats when looking at the linear equation $dX/dt = AX$.

Regarding the eigenvalue test for linear systems, let us recall, informally, the basic ideas.

The general solution of $dX/dt = AX$, assuming[9] distinct eigenvalues λ_i for A, can be written as:

$$X(t) = \sum_{i=1}^{n} c_i \, e^{\lambda_i t} v_i$$

where for each i, $Av_i = \lambda_i v_i$ (an eigenvalue/eigenvector pair) and the c_i are constants (that can be fit to initial conditions).

It is not surprising that eigen-pairs appear: if $X(t) = e^{\lambda t} v$ is solution, then $\lambda e^{\lambda t} v = dX/dt = Ae^{\lambda t}v$, which implies (divide by $e^{\lambda t}$) that $Av = \lambda v$.

---

[8]Things get very technical and difficult if A has eigenvalues with exactly zero real part. The field of mathematics called Center Manifold Theory studies that problem.
[9]If there are repeated eigenvalues, one must fine-tune a bit: it is necessary to replace some terms $c_i \, e^{\lambda_i t} v_i$ by $c_i \, t \, e^{\lambda_i t} v_i$ (or higher powers of t) and to consider "generalized eigenvectors."

We also recall that everything works in the same way even if some eigenvalues are complex, though it is more informative to express things in alternative real form (using Euler's formula).

To summarize:

- Real eigenvalues λ correspond[10] to terms in solutions that involve real exponentials $e^{\lambda t}$, which can only approach zero as $t \to +\infty$ if $\lambda < 0$.

- Non-real complex eigenvalues $\lambda = a + ib$ are associated to oscillations. They correspond[11] to terms in solutions that involve complex exponentials $e^{\lambda t}$. Since one has the general formula $e^{\lambda t} = e^{at+ibt} = e^{at}(\cos bt + i \sin bt)$, solutions, when re-written in real-only form, contain terms of the form $e^{at} \cos bt$ and $e^{at} \sin bt$, and therefore converge to zero (with decaying oscillations of "period" $2\pi/b$) provided that $a < 0$, that is to say, that the real part of λ is negative. Another way to see this if to notice that asking that $e^{\lambda t} \to 0$ is the same as requiring that the magnitude $\left| e^{\lambda t} \right| \to 0$. Since $\left| e^{\lambda t} \right| = e^{at}\sqrt{(\cos bt)^2 + (\sin bt)^2} = e^{at}$, we see once again that $a < 0$ is the condition needed in order to insure that $e^{\lambda t} \to 0$

Special Case: 2 by 2 Matrices

In the case $n = 2$, it is easy to check directly if $dX/dt = AX$ is stable, without having to actually compute the eigenvalues. Suppose that

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

and remember that

$$\text{trace } A = a_{11} + a_{22}, \quad \det A = a_{11}a_{22} - a_{12}a_{21}.$$

Then:

$$\boxed{\textit{stability is equivalent to: } \text{trace } A < 0 \textit{ and } \det A > 0.}$$

(Proof: the characteristic polynomial is $\lambda^2 + b\lambda + c$ where $c = \det A$ and $b = -\text{trace } A$. Both roots have negative real part if

$$\text{(complex case)} \quad b^2 - 4c < 0 \text{ and } b > 0$$

or

$$\text{(real case)} \quad b^2 - 4c \geq 0 \text{ and } -b \pm \sqrt{b^2 - 4c} < 0$$

and the last condition is equivalent to $\sqrt{b^2 - 4c} < b$, i.e. $b > 0$ and $b^2 > b^2 - 4c$, i.e. $b > 0$ and $c > 0$.)

Moreover, solutions are oscillatory (complex eigenvalues) if $(\text{trace } A)^2 < 4 \det A$, and exponential (real eigenvalues) otherwise. We come back to this later (trace/determinant plane).

(If you are interested: for higher dimensions ($n>2$), one can also check stability without computing eigenvalues, although the conditions are more complicated; google *Routh-Hurwitz Theorem*.)

---

[10]To be precise, if there are repeated eigenvalues, one may need to also consider terms of the slightly more complicated form "$t^k e^{\lambda t}$" but the reasoning is exactly the same in that case.

[11]For complex repeated eigenvalues, one may need to consider terms $t^k e^{\lambda t}$.

2.2.4 Chemostat: Local Stability

Let us assume that the positive equilibrium $\bar{X}_2$ exists, that is:

$$\alpha_1 > 1 \quad \text{and} \quad \beta = \alpha_2(\alpha_1 - 1) - 1 > 0 \,.$$

In that case, the Jacobian is:

$$A = F'(\bar{X}_2) = \begin{bmatrix} 0 & \beta(\alpha_1 - 1) \\ -\dfrac{1}{\alpha_1} & -\dfrac{\beta(\alpha_1 - 1) + \alpha_1}{\alpha_1} \end{bmatrix}$$

where we used the shorthand: $\beta = \alpha_2(\alpha_1 - 1) - 1$.

The trace of this matrix A is negative (because $\beta > 0$, $\alpha_1 - 1 > 0$, $\alpha_1 > 0$), and the determinant is positive:

$$\alpha_1 - 1 > 0 \quad \text{and} \quad \beta > 0 \quad \Rightarrow \quad \frac{\beta(\alpha_1 - 1)}{\alpha_1} > 0 \,.$$

So we conclude (local) stability of the positive equilibrium.

So, at least, if the initial the concentration $X(0)$ is close to $\bar{X}_2$, then $X(t) \to \bar{X}_2$ as $t \to \infty$.
(We later see that global convergence holds as well.)

What about the other equilibrium, $\bar{X}_1 = (0, \alpha_2)$? We compute the Jacobian:

$$A = F'(\bar{X}_1) = \left(\begin{array}{cc} \alpha_1 \dfrac{C}{1+C} - 1 & \dfrac{\alpha_1 N}{(1+C)^2} \\ -\dfrac{C}{1+C} & -\dfrac{N}{(1+C)^2} - 1 \end{array} \right) \Bigg|_{N=0, C=\alpha_2} = \left(\begin{array}{cc} \alpha_1 \dfrac{\alpha_2}{1+\alpha_2} - 1 & 0 \\ -\dfrac{\alpha_2}{1+\alpha_2} & -1 \end{array} \right)$$

and thus see that its determinant is:

$$1 - \alpha_1 \frac{\alpha_2}{1+\alpha_2} = \frac{1 + \alpha_2 - \alpha_1 \alpha_2}{1+\alpha_2} = \frac{1 + \alpha_2(1 - \alpha_1)}{1+\alpha_2} = -\frac{\beta}{1+\alpha_2} < 0$$

and therefore the steady state $\bar{X}_1$ is *unstable*.

It turns out that the point $\bar{X}_1$ is a saddle: small perturbations, where $N(0) > 0$, will tend away from $\bar{X}_1$. (Intuitively, if even a small amount of bacteria is initially present, growth will occur. As it turns out, the growth is so that the other equilibrium $\bar{X}_1$ is approached.)

2.3   More Modeling Examples

2.3.1   Effect of Drug on Cells in an Organ

A modification of the chemostat model can be used as a simple model of how
a drug in the blood (e.g. a chemotherapy agent) affects a cells in a certain organ
(or more specifically, a subset of cells, such as cancer cells).

Now, "C_0" represents the concentration of the drug in the blood flowing in,
and V is the volume of blood in the organ, or, more precisely,
the volume of blood in the region where the cells being treated (e.g., a tumor).

V = volume of blood
$F = F_{\text{in}}, F_{\text{out}}$ are the blood flows
$N(t)$ = number of cells (assumed equal in mass) exposed to drug
$C_0, C(t)$ = drug concentrations
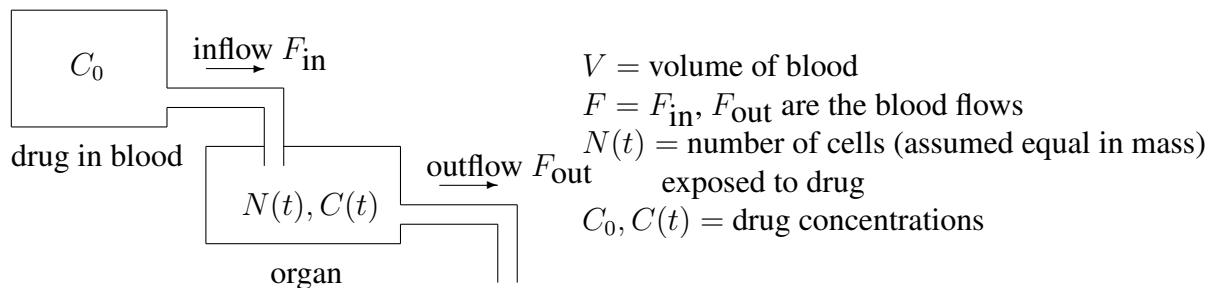
In drug infusion models, if a pump delivers the drug at a certain concentration,
the actual C_0 would account for the dilution rate when injected into the blood.

We assume that things are "well-mixed" although more realistic models use the fact
that drugs may only affect e.g. the outside layers of a tumor.

The flow F represents blood brought into the organ through an artery, and the blood coming out.

The key differences with the chemostat are:

- the cells in question reproduce at a rate that is, in principle, independent of the drug,

- but the drug has a negative effect on the growth, a "kill rate" that we model by some function $K(C)$, and

- the outflow contains only (unused) drug, and not any cells.

If we assume that cells reproduce exponentially and the drug is consumed at a rate proportional to the kill rate $K(C)N$, we are led to:

$$
\begin{aligned}
\frac{dN}{dt} &= -K(C)N + kN \\
\frac{dC}{dt} &= -\alpha K(C)N - \frac{CF_{\text{out}}}{V} + \frac{C_0 F_{\text{in}}}{V} \, .
\end{aligned}
$$

2.3.2   Compartmental Models
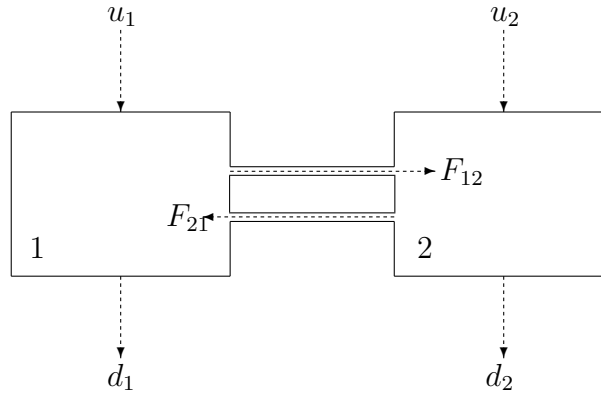
Compartmental models are very common in pharmacology and many other biochemical applications.

They are used to account for different behaviors in different tissues.

In the simplest case, there are two compartments, such as an organ and the blood in circulation.

We model the two-compartment case now (the general case is similar).

We use two variables x_1, x_2, for the concentrations (mass/vol) of a substance
(such as a drug, a hormone, a metabolite, a protein, or some other chemical) in each compartment,
and m_1, m_2 for the respective masses.

The flow (vol/sec) from compartment i to compartment j is denoted by F_{ij}.

When the substance happens to be in compartment i, a fraction $d_i \Delta t$ of its mass, degrades, or is consumed, in any small interval of time Δt,

Sometimes, there may also be an external source of the substance, being externally injected; in that case, we let u_i denote the inflow (mass/sec) into compartment i.

On a small interval Δt, the increase (or decrease, if the number is negative) of the mass in the first compartment is:

$$m_1(t + \Delta t) - m_1(t) \ = \ -F_{12} x_1 \Delta t + F_{21} x_2 \Delta t - d_1 m_1 \Delta t + u_1 \Delta t \,.$$

(For example, the mass flowing in from compartment 1 to compartment 2 is computed as:

$$\text{flow} \times \text{concentration in } 1 \times \text{time} \ = \ \frac{\text{vol}}{\text{time}} \times \frac{\text{mass}}{\text{vol}} \times \text{time} \ .)$$

Similarly, we have an equation of m_2. We divide by Δt and take limits as $\tau \to 0$, leading to the following system of two linear differential equations:

$$\frac{dm_1}{dt} \ = \ -F_{12} m_1/V_1 + F_{21} m_2/V_2 - d_1 m_1 + u_1$$

$$\frac{dm_2}{dt} \ = \ F_{12} m_1/V_1 - F_{21} m_2/V_2 - d_2 m_2 + u_2$$

(we used that $x_i = m_i/V_i$). So, for the concentrations $x_i = m_i/V_i$, we have:

$$\frac{dx_1}{dt} \ = \ -\frac{F_{12}}{V_1} x_1 + \frac{F_{21}}{V_1} x_2 - d_1 x_1 + \frac{u_1}{V_1}$$

$$\frac{dx_2}{dt} \ = \ \frac{F_{12}}{V_2} x_1 - \frac{F_{21}}{V_2} x_2 - d_2 x_2 + \frac{u_2}{V_2}$$

2.4 Geometric Analysis: Vector Fields, Phase Planes

2.4.1 Review: Vector Fields

One interprets $\frac{dX}{dt} = F(X)$ as a "flow" in $\mathbb{R}^n$: at each position X, $F(X)$ is a vector that indicates in which direction to move (and its magnitude says at what speed).
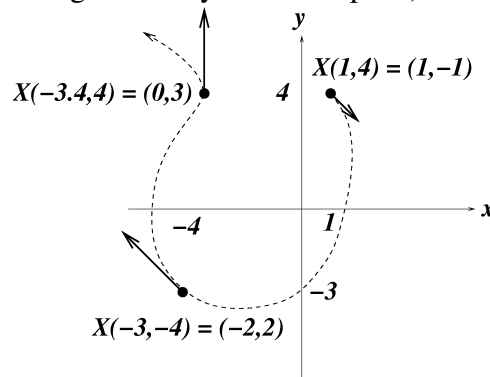
("go with the flow" or "follow directions").

We draw pictures in two dimensions, but this geometric interpretation is valid in any dimension.

"Zooming in" at steady states[12] $\bar{X}$ amounts to looking at the linearization $F(X) \approx AX$, where $A = $ Jacobian $F'(\bar{X})$ evaluated at this equilibrium.

You should work-out some phase planes using JOde or some other package.

2.4.2 Review: Linear Phase Planes

Cases of distinct real and nonzero[13] eigenvalues $\lambda_1 \neq \lambda_2$:

1. both λ_1, λ_2 are negative: *sink* (stable node)

   all trajectories approach the origin, tangent to the direction of eigenvectors corresponding to the eigenvalue which is closer to zero.

2. both λ_1, λ_2 are positive: *source* (unstable node)

   all trajectories go away from the origin, tangent to the direction of eigenvectors corresponding to the eigenvalue which is closer to zero.

3. λ_1, λ_2 have opposite signs: *saddle*

Cases of complex eigenvalues λ_1, λ_2, i.e. $= a \pm ib$ ($b \neq 0$):

1. $a = 0$: *center*

---

[12]Zooming into points that are not equilibria is not interesting; a theorem called the "flow box theorem" says (for a vector field defined by differentiable funcions) that the flow picture near a point $\bar{X}$ that is not an equilibrium is quite "boring" as it consists essentially of a bundle of parallel lines.

[13]The cases when one or both eigenvalues are zero, or are both nonzero but equal, can be also analyzed, but they are a little more complicated.

solutions[14] look like ellipses (or circles);

to decide if they more clockwise or counterclockwise, just pick one point in the plane and see which direction Ax points to;

the plots of $x(t)$ and $y(t)$ vs. time look roughly like a graph of sine or cosine.

2. $a < 0$: *spiral sink* (stable spiral)

trajectories go toward the origin while spiraling around it, and direction can be figured out as above;

the plots of $x(t)$ and $y(t)$ vs. time look roughly like the graph of a sine or cosine that is dying out (damped oscillation).

3. $a > 0$: *spiral source* (unstable spiral)

trajectories go away from the origin while spiraling around it, and direction can be figured out as above;

the plots of $x(t)$ and $y(t)$ vs. time look roughly like the graph of a sine or cosine that that is exploding (increasing oscillation).
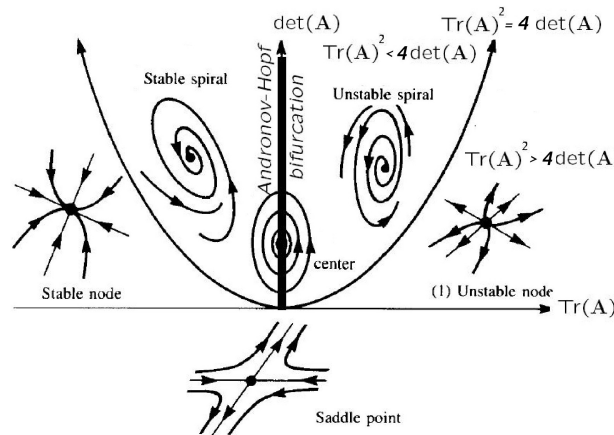
Trace/Determinant Plane

We next compute the type of the local equilibria for the chemostat example, assuming that $\alpha_1 > 1$ and $\alpha_2(\alpha_1 - 1) - 1 > 0$ (so $\bar{X}_2$ is positive).

Recall that the we had computed the Jacobian at the positive equilibrium $\bar{X}_2 = \left(\alpha_1 \left(\alpha_2 - \frac{1}{\alpha_1 - 1} \right), \frac{1}{\alpha_1 - 1} \right)$:

$$A = F'(\bar{X}_2) = \begin{bmatrix} 0 & \beta(\alpha_1 - 1) \\ -\dfrac{1}{\alpha_1} & -\dfrac{\beta(\alpha_1 - 1) + \alpha_1}{\alpha_1} \end{bmatrix}$$

where we used the shorthand: $\beta = \alpha_2(\alpha_1 - 1) - 1$.

We already saw that the trace is negative. Note that:

$$\mathrm{tr}(A) = -1 - \Delta, \quad \text{where } \Delta = \det(A) = \frac{\beta(\alpha_1 - 1)}{\alpha_1} > 0$$

and therefore $\mathrm{tr}^2 - 4\det = 1 + 2\Delta + \Delta^2 - 4\Delta = (1 - \Delta)^2 > 0$, so the point $\bar{X}_2$ is a *stable node*.[15]

Show as an exercise that $\bar{X}_1$ is a saddle.

---

[14]Centers are highly "non-robust" in a way that we will discuss later, so they rarely appear in realistic biological models.
[15]If $\Delta \neq 1$; otherwise there are repeated real eigenvalues; we still have stability, but we'll ignore that very special case.

2.4.3   Nullclines

Linearization helps understand the "local" picture of flows.[16]

It is much harder to get *global* information, telling us how these local pictures fit together ("connecting the dots" so to speak).

One useful technique when drawing global pictures is that of *nullclines*.

The x_i-nullcline (if the variables are called $x_1, x_2, \ldots$) is the set where $\frac{dx_i}{dt} = 0$.
This set may be the union of several components (curves and lines), or just one such component.[17]

The intersections between the nullclines are the steady states. This is because each nullcline is the set where $dx_1/dt = 0, dx_2/dt = 0, \ldots$, so intersecting gives points at which *all* $dx_i/dt = 0$, that is to say $F(X) = 0$ which is the definition of steady states.

As an example, let us take the chemostat, for which the vector field is $F(X) = \begin{pmatrix} f(N, C) \\ g(N, C) \end{pmatrix}$, where:

$$
\begin{aligned}
f(N, C) &= \alpha_1 \frac{C}{1 + C} N - N \\
g(N, C) &= -\frac{C}{1 + C} N - C + \alpha_2 .
\end{aligned}
$$

The N-nullcline is the set where $dN/dt = 0$, that is, where $\alpha_1 \frac{C}{1+C} N - N = 0$.
Since we can factor this as $N(\alpha_1 \frac{C}{1+C} - 1) = 0$, we see that:

the N-nullcline is the union of a horizontal and a vertical line: $\quad C = \dfrac{1}{\alpha_1 - 1} \quad$ and $\quad N = 0$.

On this set, the arrows are vertical, because $dN/dt = 0$ (no movement in N direction).

The C-nullcline is obtained by setting $-\frac{C}{1+C} N - C + \alpha_2 = 0$.
We can describe a curve in any way we want; in this case, it is a little simpler to solve $N = N(C)$ than $C = C(N)$:

the C-nullcline is the curve: $\quad N = (\alpha_2 - C)\dfrac{1 + C}{C} = -1 - C + \dfrac{\alpha_2}{C} + \alpha_2$.

On this set, the arrows are parallel to the N-axis, because $dC/dt = 0$ (no movement in C direction).

To plot, note that $N(\alpha_2) = 0$ and $N(C)$ is a decreasing function of C and goes to $+\infty$ as $C \searrow 0$, and then obtain $C = C(N)$ by flipping along the main diagonal (dotted and dashed curves in the graph, respectively):

---

[16] Actually, linearization is sometimes not sufficient even for local analysis. Think of $dx/dt = x^3$ and $dx/dt = -x^3$, which have the same linearization ($dx/dt = 0$) but very different local pictures at zero. The area of mathematics called "Center manifold theory" deals with such very special situations, where eigenvalues may be zero or more generally have zero real part.

[17] Some authors like to call each component a "nullcline" but I prefer to say that the nullcline is one object, which may have more than one component. It is just a question of semantics.
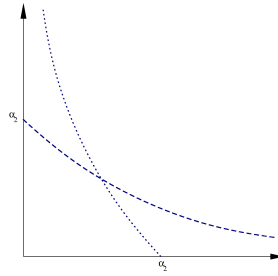
In summary, the nullclines look as follows:
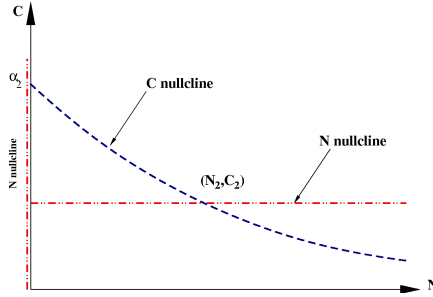
Assuming that $\alpha_1 > 1$ and $\alpha_2 > 1/(\alpha_1 - 1)$, so that a positive steady-state exists, we have the two intersections: $(0, \alpha_2)$ (saddle) and $\left(\alpha_1 \left(\alpha_2 - \frac{1}{\alpha_1 - 1} \right), \frac{1}{\alpha_1 - 1} \right)$ (stable node).

To decide whether the arrows point up or down on the N-nullcline, we need to look at dC/dt.

On the line $N = 0$ we have:

$$\frac{dC}{dt} = -\frac{C}{1 + C} N - C + \alpha_2 = -C + \alpha_2 \begin{cases} > 0 & \text{if } C < \alpha_2 \\ < 0 & \text{if } C > \alpha_2 \end{cases}$$

so the arrows point up if $C < \alpha_2$ and down otherwise. On the line $C = \frac{1}{\alpha_1 - 1}$:

$$\frac{dC}{dt} = -\frac{C}{1 + C} N - C + \alpha_2 = \alpha_2 - \frac{1}{\alpha_1 - 1} - \frac{N}{\alpha_1} \begin{cases} > 0 & \text{if } N < \alpha_1 \left(\alpha_2 - \frac{1}{\alpha_1 - 1} \right) \\ < 0 & \text{if } N > \alpha_1 \left(\alpha_2 - \frac{1}{\alpha_1 - 1} \right) \end{cases}$$

so the arrow points up if $N < \alpha_1 \left(\alpha_2 - \frac{1}{\alpha_1 - 1} \right)$ and down otherwise.

To decide whether the arrows point right or left (sign of dN/dt) on the C-nullcline, we look at:

$$\frac{dN}{dt} = N \left(\alpha_1 \frac{C}{1 + C} - 1 \right) \begin{cases} > 0 & \text{if } C > \dfrac{1}{\alpha_1 - 1} \\ < 0 & \text{if } C < \dfrac{1}{\alpha_1 - 1} \end{cases}$$

(since $N \geq 0$, the sign of the expression is the same as the sign of $\alpha_1 \frac{C}{1+C} - 1$).

A shortcut for determining directions on nullclines

Observe that directions cannot change in any segment (in-between intersections with the other nullcline), since a change of direction would means that the other derivative is zero (and theerfore that we must cross the other nullcline).

So, we may simply pick *any* point in such a segment to determine the direction.

For example, for the two components of the N-nullcline, we have:

(1) on the line $N = 0$: $\frac{dC}{dt} = -C + \alpha_2$, so > 0 at $C = 0$ and < 0 as $C \to +\infty$;

(2) on line $C = \frac{1}{\alpha_1 - 1}$: $\frac{dC}{dt} = \alpha_2 - \frac{1}{\alpha_1 - 1} - \frac{N}{\alpha_1}$, so > 0 at $N = 0$ (because $\alpha_2 - \frac{1}{\alpha_1 - 1} > 0$) and < 0 as $N \to +\infty$

On the C-nullcline, $\frac{dN}{dt} = N\left(\alpha_1 \frac{C}{1+C} - 1\right)$ is > 0 as $C \to \alpha_2$ (because $\alpha_1 \frac{\alpha_2}{1+\alpha_2} - 1 > 0$) and < 0 at $C = 0$.

This information is enough to determine the signs on each segment.

We have, therefore, this picture:
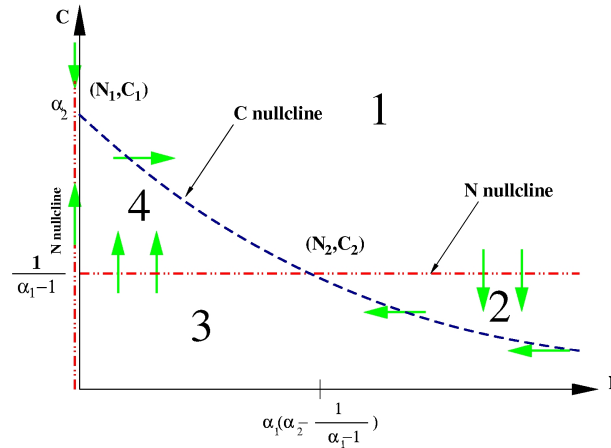
What about the direction of the vector field elsewhere, not just on nullclines?

The key observation is that the *only way* that arrows can "reverse direction" is by crossing a nullcline.

For example, if dx_1/dt is positive at some point A, and it is negative at some other point B, then A and B must be on opposite sides of the x_1 nullcline. The reason is that, were we to trace a path between A and B (any path, not necessarily a solution of the system), the derivative dx_1/dt at the points in the path varies continuously[18] and therefore (intermediate value theorem) there must be a point in this path where $dx_1/dt = 0$.

In summary: if we look at regions demarcated by the nullclines[19] then the orientations of arrows remain the same in each such region.

For example, for the chemostat, we have 4 regions, as shown in the figure.

In region 1, $dN/dt > 0$ and $dC/dt < 0$, since these are the values in the boundaries of the region. Therefore the flow is "Southeast" (↘) in that region. Similarly for the other three regions.

We indicate this information in the phase plane:

---

[18]assuming that the vector field is continuous

[19]the "connected components" of the complement of the nullclines, think of them as the "territories" separated by the nullclines
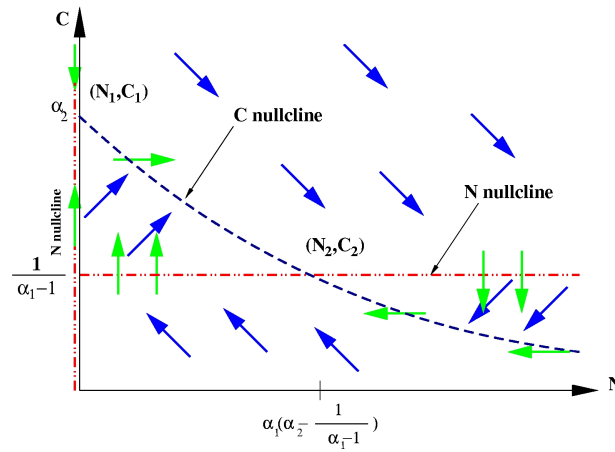
Note that the arrows are just "icons" intended to indicate if the flow is generally "SE" ($dN/dt > 0$ and $dC/dt < 0$), "NE," etc, but the actual numerical slopes will vary (for example, near the nullclines, the arrows must become either horizontal or vertical).

2.4.4 Global Behavior

We already know that trajectories that start *near* the positive steady state $\bar{X}_2$ converge to it (local stability)

and that most trajectories that start near $\bar{X}_1$ go away from it (instability).

(Still assuming, obviously, that the parameters have been chosen in such a way that the positive steady state exists.)

Let us now sketch a proof that, in fact, *every* trajectory converges to $\bar{X}_2$ (with the exception only of those trajectories that start with $N(0) = 0$).

The practical consequences of this "global attraction" result are that, no matter what the initial conditions, the chemostat will settle into the steady state $\bar{X}_2$.

It is helpful to consider the following line:

$$(L) \qquad N + \alpha_1 C - \alpha_1 \alpha_2 = 0$$

which passes through the points $\bar{X}_1 = (0, \alpha_2)$ and $\bar{X}_2 = \left(\alpha_1 \left(\alpha_2 - \frac{1}{\alpha_1 - 1} \right), \frac{1}{\alpha_1 - 1} \right)$.

Note that $(\alpha_1 \alpha_2, 0)$ is also in this line.

The picture is as follows[20] where the arrows are obtained from the flow direction, as shown earlier.

---

[20]you may try as an exercise to show that the C-nullcline is concave up, so it must intersect L at just two points, as shown
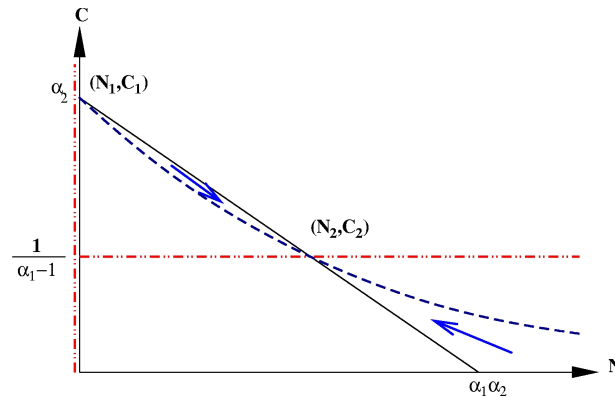
We claim that this line is *invariant*, that is, solutions that start in L must remain in L. Even more interesting, all trajectories (except those that start with $N(0) = 0$) converge to L.

For any trajectory, consider the following function:

$$z(t) = N(t) + \alpha_1 C(t) - \alpha_1 \alpha_2$$

and observe that

$$z' = N' + \alpha_1 C' = \alpha_1 \frac{C}{1+C} N - N - \alpha_1 \left(\frac{C}{1+C} N - C + \alpha_2 \right) = -z$$

which implies that $z(t) = z(0)e^{-t}$. Therefore, $z(t) = 0$ for all $t > 0$, if $z(0) = 0$ (invariance), and in general $z(t) \to 0$ as $t \to +\infty$ (solutions approach L).

Moreover, points in the line $N + \alpha_1 C - \alpha_1 \alpha_2 = m$ are close to points in L if m is near zero.

Since L is invariant and there are no steady states in L except $\bar{X}_1$ and $\bar{X}_2$, the open segment from $\bar{X}_1$ to $\bar{X}_2$ is a trajectory that "connects" the unstable state $\bar{X}_1$ to the stable state $\bar{X}_2$. Such a trajectory is called a *heteroclinic connection*.[21]

Now, we know that all trajectories approach L, and cannot cross L (no trajectories can ever cross, by uniqueness of solutions, as seen in your ODE class).

Suppose that a trajectory starts, and hence remains, on top of L (the argument is similar if remains under L), and with $N(0) > 0$.

Since the trajectory gets closer and closer to L, and must stay in the first quadrant (why?), it will either converge to $\bar{X}_2$ "from the NW" or it will eventually enter the region with the "NW arrow" – at which point it must have turned and start moving towards $\bar{X}_2$. In summary, every trajectory converges.

---
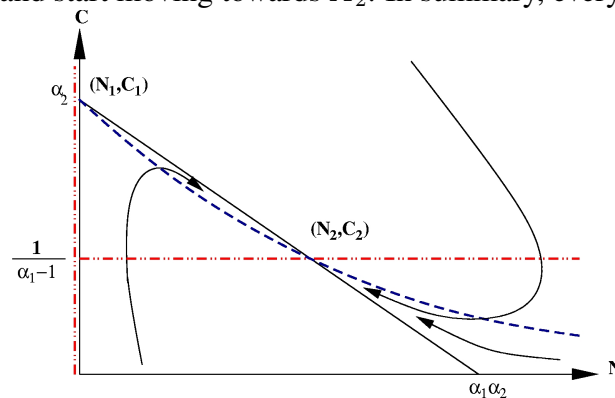
[21]Exercise: check eigenvectors at $\bar{X}_1$ and $\bar{X}_2$ to see that L matches the linearized eigen-directions.

2.5   Epidemiology: SIRS Model

The modeling of infectious diseases and their spread is an important part of mathematical biology, part of the field of mathematical epidemiology.

Modeling is an important tool for gauging the impact of different vaccination programs on the control or eradication of diseases.

We will only study here a simple ODE model, which does not take into account age structure nor geographical distribution. More sophisticated models can be based on compartmental systems, with compartments corresponding to different age groups, partial differential equations, where independent variables specify location, and so on, but the simple ODE model already brings up many of the fundamental ideas.

The classical work on epidemics dates back to Kermack and McKendrick, in 1927. We will study their SIR and SIRS models without "vital dynamics" (births and deaths).

To explain the model, let us think of a flu epidemic, but the ideas are very general.

In the population, there will be a group of people who are *Susceptible* to being passed on the virus by the *Infected* individuals.

At some point, the infected individuals get so sick that they have to stay home, and become part of the *Removed* group. Once that they recover, they still cannot infect others, nor can they be infected since they developed immunity.

The numbers of individuals in the three classes will be denoted by S, I, and R respectively, and hence the name "SIR" model.

Depending on the time-scale of interest for analysis, one may also allow for the fact that individuals in the Removed group may eventually return to the Susceptible population, which would happen if immunity is only temporary. This is the "$SIRS$" model (the last S to indicate flow from R to S), which we will study next.

We assume that these numbers are all functions of time t, and that the numbers can be modeled as real numbers. (Non-integers make no sense for populations, but it is a mathematical convenience. Or, if one studies probabilistic instead of deterministic models, these numbers represent expected values of random variables, which can easily be non-integers.)

The basic modeling assumption is that the number of new infectives $I(t+\Delta t)-I(t)$ in a small interval of time $[t, t + \Delta t]$ is proportional to the product $S(t)I(t)\,\Delta t$.

Let us try to justify intuitively why it makes sense. (As usual, experimentation and fitting to data should determine if this is a good assumption. In fact, alternative models have been proposed as well.)

Suppose that transmission of the disease can happen only if a susceptible and infective are very close to each other, for instance by direct contact, sneezing, etc.

We suppose that there is some region around a given susceptible individual, so that he can only get infected if an infective enters that region:
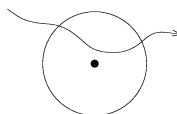
We assume that, for each infective individual, there is a probability $p = \beta \Delta t$ that this infective will

happen to pass through this region in the time interval $[t, t + \Delta t]$, where β is some positive constant that depends on the size of the region, how fast the infectives are moving, etc. (Think of the infective traveling at a fixed speed: in twice the length of time, there is twice the chance that it will pass by this region.) We take $\Delta t \ll 0$, so also $p \ll 0$.

The probability that this particular infective will *not* enter the region is $1 - p$, and, assuming independence, the probability than *no infective* enters is $(1 - p)^I$.

So the probability that *some* infective comes close to our susceptible is, using a binomial expansion: $1 - (1 - p)^I \approx 1 - (1 - pI + \binom{I}{2}p^2 + \ldots) \approx pI$ since $p \ll 1$.

Thus, we can say that a particular susceptible has a probability pI of being infected. Since there are S of them, we may assume, if S is large, that the total number infected will be $S \times pI$.

We conclude that the number of new infections is:

$$I(t + \Delta t) - I(t) = pSI = \beta SI \, \Delta t$$

and dividing by Δt and taking limits, we have a term βSI in $\frac{dI}{dt}$, and similarly a term $-\beta SI$ in $\frac{dS}{dt}$.

This is called a *mass action kinetics* assumption, and is also used when writing elementary chemical reactions. In chemical reaction theory, one derives this mass action formula using "collision theory" among particles (for instance, molecules), taking into account temperature (which affects how fast particles are moving), shapes, etc.

We also have to model infectives being removed: it is reasonable to assume that a certain fraction of them is removed per unit of time, giving terms νI, for some constant ν.

Similarly, there are terms γR for the "flow" of removeds back into the susceptible population.
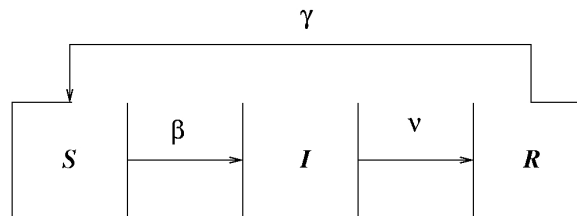
The figure is a little misleading: this is *not* a compartmental system, in which the flow from S to I is just proportional to S. For example, when $I = 0$, no one gets infected; hence the product term in the equations:

$$\begin{aligned}
\frac{dS}{dt} &= -\beta SI &&+ \gamma R \\
\frac{dI}{dt} &= \beta SI - \nu I \\
\frac{dR}{dt} &= \nu I - \gamma R
\end{aligned}$$

(There are many variations possible; here are some. In a model with vital dynamics, one also adds birth and death rates to this model. Another one: a vaccine is given to a certain percentage of the susceptibles, at a given rate, causing the vaccinated individuals to become "removed". Yet another one: there is a type of mosquito that makes people infected.)

2.5.1 Analysis of Equations

Let $N = S(t) + I(t) + R(t)$. Since $dN/dt = 0$, N is constant, the total size of the population.

Therefore, even though we are interested in a system of three equations, this *conservation law* allows us to eliminate one equation, for example, using $R = N - S - I$.

We are led to the study of the following two dimensional system:

$$\begin{aligned} \frac{dS}{dt} &= -\beta SI + \gamma(N - S - I) \\ \frac{dI}{dt} &= \beta SI - \nu I \end{aligned}$$

I-nullcline: union of lines $I = 0$ and $S = \nu/\beta$.

S-nullcline: curve $I = \frac{\gamma(N-S)}{S\beta+\gamma}$.

The steady states are

$$\bar{X}_1 = (N, 0) \quad \text{and} \quad \bar{X}_2 = \left(\frac{\nu}{\beta}, \frac{\gamma(N - \frac{\nu}{\beta})}{\nu + \gamma} \right),$$

where $\bar{X}_2$ only makes physical sense if the following condition is satisfied:

$$\boxed{\text{``}\sigma\text{''} \quad \text{or} \quad \text{``}R_0\text{''} \;=\; N\beta/\nu > 1}$$

some estimated values of R_0 are as follows (Wikipedia Oct 2014):

Disease	Transmission	R_0
Measles	Airborne	12–18
Pertussis	Airborne droplet	12–17
Diphtheria	Saliva	6–7
Smallpox	Airborne droplet	5–7
Polio	Fecal-oral route	5–7
Rubella	Airborne droplet	5–7
Mumps	Airborne droplet	4–7
HIV/AIDS	Sexual contact	2–5
SARS	Airborne droplet	2–5[2]
Influenza (1918 pandemic strain)	Airborne droplet	2–3[3]
Ebola (2014 Ebola outbreak)	Bodily fluids	1-2 [4]

For example, if $N = 2$, $\beta = 1$, $\nu = 1$, and $\gamma = 1$, the I-nullcline is the union of $I=0$ and $S=1$, the S-nullcline is given by $I = \frac{(2-S)}{S+1}$, and the equilibria are at $(2, 0)$ and $(1, 1/2)$
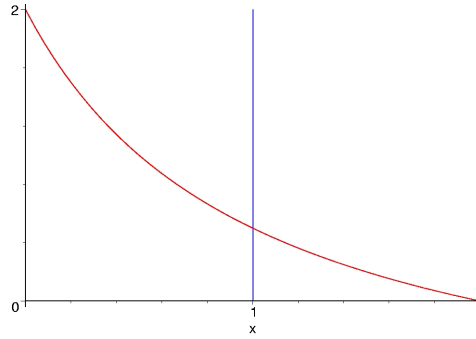
The Jacobian is, at any point:

$$\begin{bmatrix} -I\beta - \gamma & -S\beta - \gamma \\ I\beta & S\beta - \nu \end{bmatrix}$$

so the trace and determinant at $\bar{X}_1 = (N, 0)$ are, respectively:

$$-\gamma + N\beta - \nu \quad \text{and} \quad -\gamma(N\beta - \nu)$$

and thus, provided $R_0 = N\beta/\nu > 1$, we have det< 0 and hence a saddle.

At $\bar{X}_2$ we have: trace $= -I\beta - \gamma < 0$ and det $= I\beta(\nu + \gamma) > 0$, and hence this steady state is stable.

Therefore, at least for close enough initial conditions (since the analysis is local, we cannot say more), and assuming $R_0 > 1$, the number of infected individuals will approach

$$I_{\text{steady state}} = \frac{\gamma(N - \frac{\nu}{\beta})}{\nu + \gamma}.$$

2.5.2 Interpreting R_0

Let us give an intuitive interpretation of R_0.

We make the following "thought experiment":
suppose that we isolate a group of P infected individuals, and allow them to recover.

Since there are no susceptibles in our imagined experiment, $S(t) \equiv 0$, so $\frac{dI}{dt} = -\nu I$, so $I(t) = Pe^{-\nu t}$.

Suppose that the ith individual is infected for a total of d_i days, and look at the following table:

cal. days→ / Individuals	0	1	2		...	d_1	∞	
Ind. 1	X	X	X	X	X	X		$= d_1$ days
Ind. 2	X	X	X	X				$= d_2$ days
Ind. 3	X	X	X	X	X			$= d_3$ days
. . .								
Ind. P	X	X	X	X				$= d_P$ days
	$=I_0$	$=I_1$	$=I_2$	...				

It is clear that $d_1 + d_2 + \ldots = I_0 + I_1 + I_2 + \ldots$
(supposing that we count on integer days, or hours, or some other discrete time unit).

Therefore, the average number of days that individuals are infected is:

$$\frac{1}{P}\sum d_i \;\; = \;\; \frac{1}{P}\sum I_i \;\; \approx \;\; \frac{1}{P}\int_0^\infty I(t)\,dt \;\; = \;\; \int_0^\infty e^{-\nu t}\,dt \;\; = \;\; \frac{1}{\nu}.$$

On the other hand, back to the original model, what is the meaning of the term "βSI" in dI/dt?

It means that $I(\Delta t) - I(0) \approx \beta S(0)I(0)\Delta t$.

Therefore, if we start with $I(0)$ infectives, and we look at an interval of time of length $\Delta t = 1/\nu$, which we agreed represents the average time of an infection, we end up with the following number of new infectives:

$$\beta(N - I(0))I(0)/\nu \;\; \approx \;\; \beta N I(0)/\nu$$

if $I(0) \ll N$, which means that each individual, on the average, infected $(\beta N I(0)/\nu)/I(0) = R_0$ new individuals.

We conclude, from this admittedly hand-waving argument[22], that R_0 represents the *expected number infected by a single individual* (in epidemiology, the *intrinsic reproductive rate* of the disease).

2.5.3 Nullcline Analysis

For the previous example, $N = 2$, $\beta = 1$, $\nu = 1$, and $\gamma = 1$:

$$\frac{dS}{dt} \;\; = \;\; -SI + 2 - S - I$$
$$\frac{dI}{dt} \;\; = \;\; SI - I$$

with equilibria at $(2, 0)$ and $(1, 1/2)$, the I-nullcline is the union of $I=0$ and $S=1$.

When $I = 0$, $dS/dt = 2 - S$,
and on $S = 1$, $dS/dt = 1 - 2I$,
so we can find if arrows are right or left pointing.

On the S-nullcline $I = \frac{(2-S)}{S+1}$ we have

$$\frac{dI}{dt} \;\; = \;\; \frac{(S-1)(2-S)}{S+1}$$

and therefore arrows point down if $S < 1$, and up if $S \in (1, 2)$. This in turn allows us to know the general orientation (NE, etc) of the vector field.
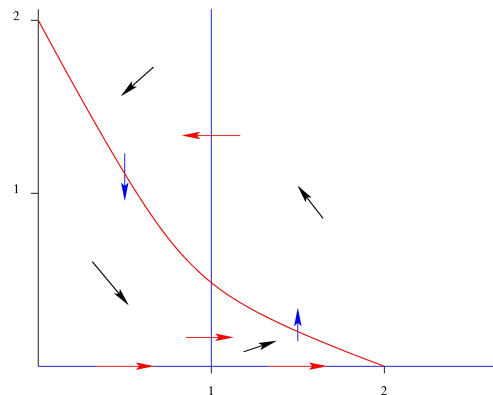
Here are computer-generated phase-planes[23] for this example as well as for a modification in which we took $\nu = 3$ (so $R_0 < 1$).

---

[22]among other things, we'd need to know that ν is large, so that Δt is small
[23]Physically, only initial conditions with $I + S \leq 2$ make sense; why?
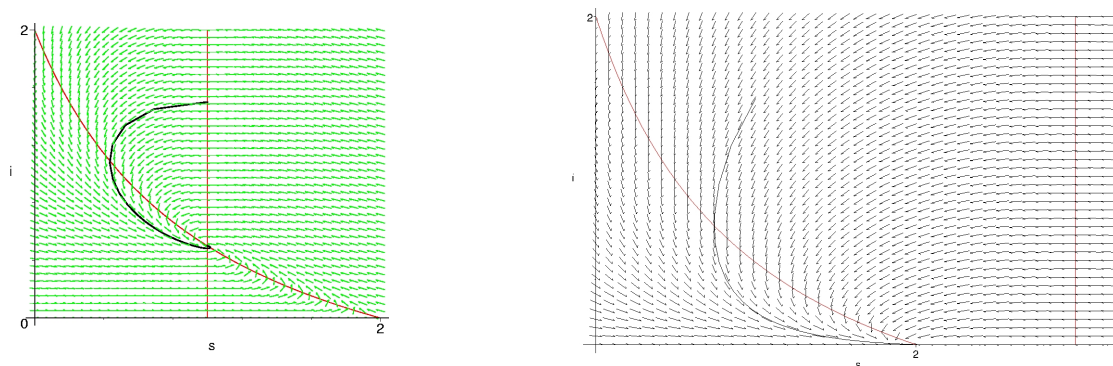
In the first case, the system settles to the positive steady state, no matter where started, as long as $I(0) > 0$.

In the second case, there is only one equilibrium, since the vertical component of the I-nullcline is at $S = 3/1 = 3$, which does not intersect the other nullcline. The disease will disappear in this case.

Example simulation ($\beta = .003$, $\nu = 1$, $\gamma = 0.5$):
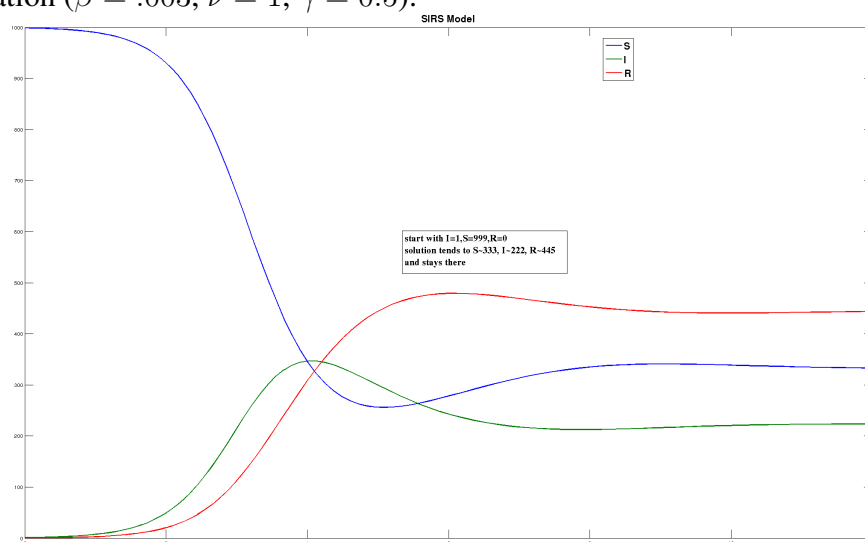
Another example simulation (same parameters, different initial):
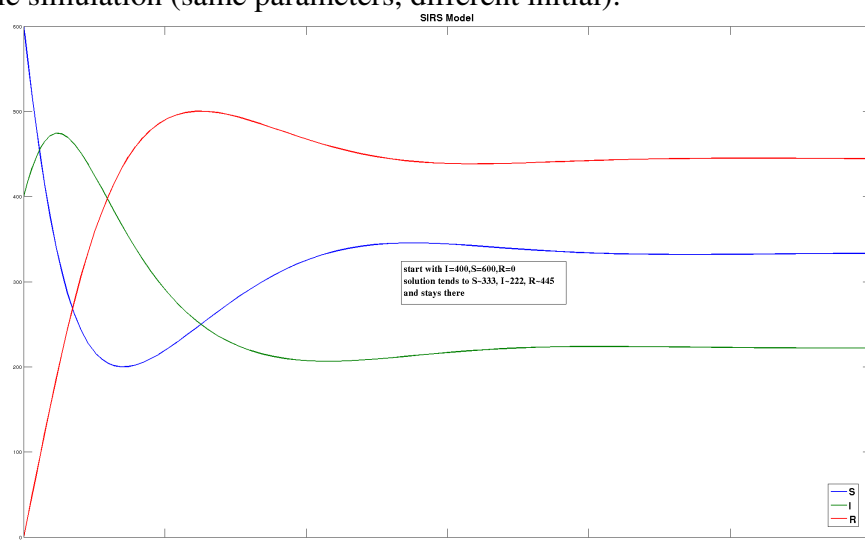
2.5.4 Immunizations

The effect of immunizations is to reduce the "threshold" N needed for a disease to take hold.

In other words, for N small, the condition $R_0 = N\beta/\nu > 1$ will fail, and no positive steady state will exist.

Vaccinations have the effect to permanently remove a certain proportion p of individuals from the population, so that, in effect, N is replaced by pN. Vaccinating just $p > 1 - \frac{1}{R_0}$ individuals gives $(1-p)R_0 < 1$, and hence suffices to eradicate a disease!

2.5.5 A Variation: STD's

Suppose that we wish to study a virus that can only be passed on by heterosexual sex. Then we should consider two separate populations, male and female. We use $\bar{S}$ to indicate the susceptible males and S for the females, and similarly for I and R.

The equations analogous to the SIRS model are:

$$
\begin{aligned}
\frac{d\bar{S}}{dt} &= -\bar{\beta}\bar{S}I && + \bar{\gamma}\bar{R} \\
\frac{d\bar{I}}{dt} &= \bar{\beta}\bar{S}I - \bar{\nu}\bar{I} \\
\frac{d\bar{R}}{dt} &= && \bar{\nu}\bar{I} - \bar{\gamma}\bar{R} \\
\frac{dS}{dt} &= -\beta S\bar{I} && + \gamma R \\
\frac{dI}{dt} &= \beta S\bar{I} - \nu I \\
\frac{dR}{dt} &= && \nu I - \gamma R \,.
\end{aligned}
$$

This model is a little difficult to study, but in many STD's (especially asymptomatic), there is no "removed" class, but instead the infecteds get back into the susceptible population. This gives:

$$
\begin{aligned}
\frac{d\bar{S}}{dt} &= -\bar{\beta}\bar{S}I && + \bar{\nu}\bar{I} \\
\frac{d\bar{I}}{dt} &= \bar{\beta}\bar{S}I - \bar{\nu}\bar{I} \\
\frac{dS}{dt} &= -\beta S\bar{I} && + \nu I \\
\frac{dI}{dt} &= \beta S\bar{I} - \nu I \,.
\end{aligned}
$$

Writing $\bar{N} = \bar{S}(t) + \bar{I}(t)$ and $N = S(t) + I(t)$ for the total numbers of males and females, and using these two conservation laws, we can just study the following set of two ODE's:

$$
\begin{aligned}
\frac{d\bar{I}}{dt} &= \bar{\beta}(\bar{N} - \bar{I})I - \bar{\nu}\bar{I} \\
\frac{dI}{dt} &= \beta(N - I)\bar{I} - \nu I \,.
\end{aligned}
$$

2.5.6 A Variation, "SIR"

Now assume no recovery from Removed to Susceptibles:

$$\frac{dS}{dt} = -\beta SI$$
$$\frac{dI}{dt} = \beta SI - \nu I$$
$$\frac{dR}{dt} = \nu I$$

a simulation ($\beta = .003$, $\nu = 1$) is shown below:
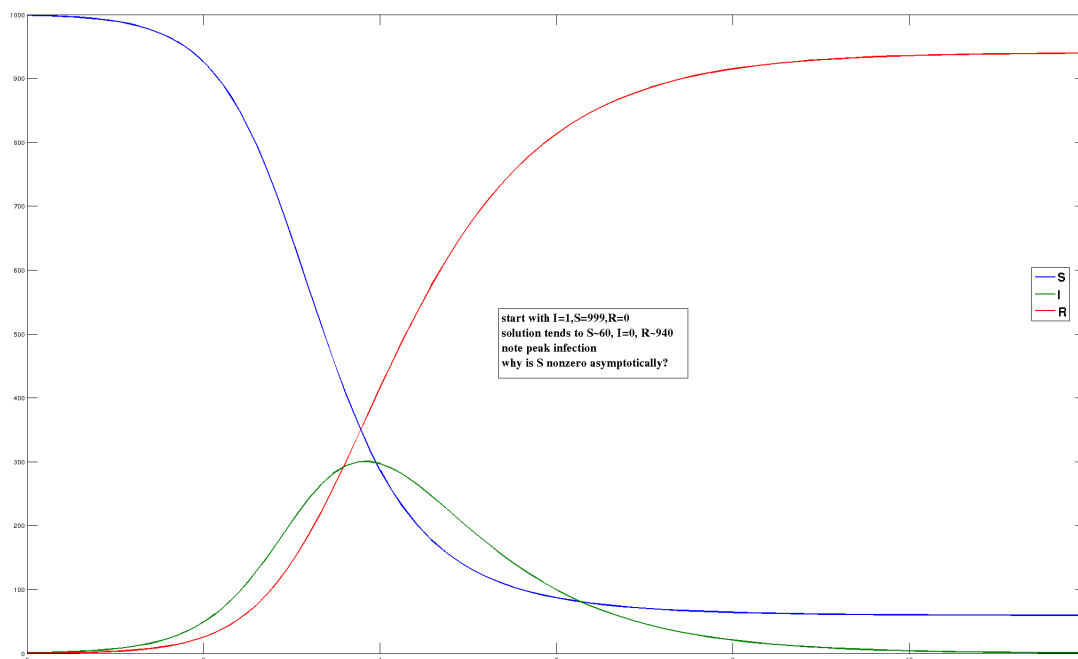
Note the interesting behavior: – analyze it as a homework problem!

In particular, think of these: why does $I(t)$ peak and go down? Why is there a residual S?

2.5.7 Another variation: "SEIR"

The SIR model has no latent stage, so it is inappropriate for certain diseases.

Thus we may add an "incubation period" as an intermediate stage between Susceptibles and Infecteds; this leads to the "SEIR" model with subpopulations: "Susceptible", "Exposed", "Infected", and "Removed". A natural set of differential equations for the SEIR model is as follows:

$$dS/dt = -\beta I(t)S(t)$$
$$dE/dt = \beta I(t)S(t) - \varepsilon E(t)$$
$$dI/dt = \varepsilon E(t) - \nu I(t)$$
$$dR/dt = \nu I(t)$$

where we may interpret ν and ε as inverses of infection and incubation periods.

As an example of SEIR models, consider the 2009-2010 flu pandemic in Istanbul (in June 2009, the World Health Organization declared A/H1N1 a pandemic).

In the paper "A susceptible-exposed-infected-removed (SEIR) model for the 2009-2010 A/H1N1 epidemic in Istanbul" by Funda Samanlioglu, Ayse Humeyra Bilge, Onder Ergonul (arXiv 1205.2497), this model is used to fit data on medical reports, dates of hospitalization, and recovery or death, from major Istanbul hospitals. The following best-fit was obtained there:
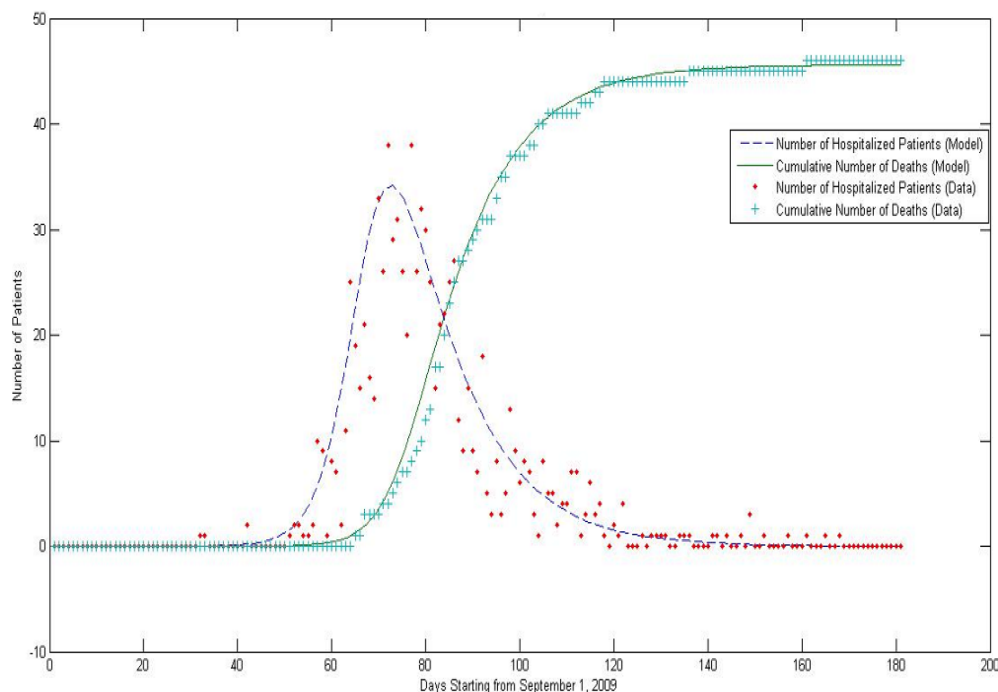
This is fairly good, qualitatively at least.

The parameters in the model are I_0, ν, ε and β, where I_0 is the percentage of people infected initially. The authors assumed for fitting that the number of fatalities was proportional to the number of removed individuals, and the number of hospitalizations proportional to the number of infections.

The parameters that they found for the best fit to the model were: $\nu = 0.09$, $I_0 = 10^{-7.4}$, $\varepsilon = 0.32$, $\beta = 0.585$ and this gave a mean-squared error of 10% and 2.6% to infections and fatalities respectively.

2.6 Chemical Kinetics

Elementary reactions (in a gas or liquid) are due to collisions of particles (molecules, atoms).

Particles move at a velocity that depends on temperature (higher temperature $\Rightarrow$ faster).

The *law of mass action* is:

> *reaction rates (at constant temperature) are proportional to products of concentrations.*

This law may be justified intuitively in various ways, for instance, using an argument like the one that we presented for disease transmission.

In chemistry, *collision theory* studies this question and justifies mass-action kinetics.

To be precise, it isn't enough for collisions to happen - the collisions have to happen in the "right way" and with enough energy for bonds to break.

For example[24] consider the following simple reaction involving a collision between two molecules: ethene (CH2=CH2) and hydrogen chloride (HCl), which results om chloroethane.

As a result of the collision between the two molecules, the double bond between the two carbons is converted into a single bond, a hydrogen atom gets attached to one of the carbons, and a chlorine atom to the other.

But the reaction can only work if the hydrogen end of the H-Cl bond approaches the carbon-carbon double bond; any other collision between the two molecules doesn't produce the product, since the two simply bounce off each other.
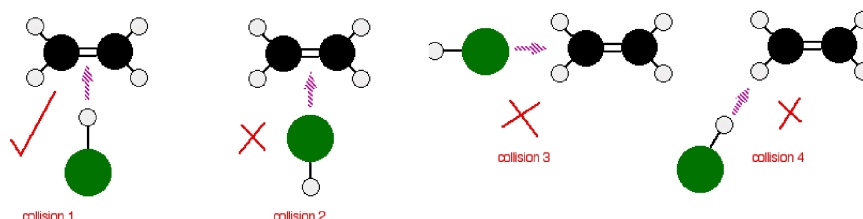
The proportionality factor (the *rate constant*) in the law of mass action accounts for temperature, probabilities of the right collision happening if the molecules are near each other, etc.

We will derive ordinary differential equations based on mass action kinetics. However, it is important to remember several points:
- If the medium is not "well mixed" then mass-action kinetics might not be valid.
- If the number of molecules is small, a probabilistic model should be used. Mass-action ODE models are only valid as averages when dealing with large numbers of particles in a small volume.
- If a catalyst is required for a reaction to take place, then doubling the concentration of a reactants does not mean that the reaction will proceed twice as fast.[25] We later study some catalytic reactions.

---

[24]discussion borrowed from http://www.chemguide.co.uk/physical/basicrates/introduction.html

[25]As an example, consider the following analog of a chemical reaction, happening in a cafeteria: $A + B \rightarrow C$, where A is the number of students, B is the food on the counters, and C represents students with a full tray walking away from the counter. If each student would be allowed to, at random times, pick food from the counters, then twice the number of students, twice the number walking away per unit of time. But if there is a person who must hand out food (our "catalyst"), then there is a maximal rate at which students will leave the counter, a rate determined by how fast the cafeteria worker can serve each student. In this case, doubling the number of students does not mean that twice the number will walking away with their food per unit of time.

2.6.1 Equations

We will use capital letters $A, B, \ldots$ for *names* of chemical substances (molecules, ions, etc), and lower-case $a, b, \ldots$ for their corresponding *concentrations*.

There is a systematic way to write down equations for chemical reactions, using a graph description of the reactions and formulas for the different kinetic terms. We discuss this systematic approach later, but for now we consider some very simple reactions, for which we can write equations directly. We simply use the mass-action principle for each separate reaction, and add up all the effects.

The simplest "reaction" is one where there is only one reactant, that can degrade[26] or decay (as in radioactive decay), or be transformed into another species, or split into several constituents.

In either case, the rate of the reaction is proportional to the concentration:
if we have twice the amount of substance X in a certain volume, then, per (small) unit of time, a certain % of the substance in this volume will disappear, which means that the concentration will diminish by that fraction.
A corresponding number of the new substances is then produced, per unit of time.

So, decay $X \xrightarrow{k} \cdot$ gives the ODE:
$$dx/dt = -kx \, ,$$

a transformation $X \xrightarrow{k} Y$ gives:

$$
\begin{aligned}
dx/dt &= -kx \\
dy/dt &= kx \, ,
\end{aligned}
$$

and a dissociation reaction $Z \xrightarrow{k} X + Y$ gives:

$$
\begin{aligned}
dx/dt &= kz \\
dy/dt &= kz \\
dz/dt &= -kz \, .
\end{aligned}
$$

A bimolecular reaction $X + Y \xrightarrow{k_+} Z$ gives:

$$
\begin{aligned}
dx/dt &= -k_+ xy \\
dy/dt &= -k_+ xy \\
dz/dt &= k_+ xy
\end{aligned}
$$

and if the reverse reaction $Z \xrightarrow{k_-} X + Y$ also takes place:

$$
\begin{aligned}
dx/dt &= -k_+ xy + k_- z \\
dy/dt &= -k_+ xy + k_- z \\
dy/dt &= k_+ xy - k_- z \, .
\end{aligned}
$$

---

[26]Of course, "degrade" is a relative concept, because the separate parts of the decaying substance should be taken account of. However, if these parts are not active in any further reactions, one ignores them and simply thinks of the reactant as disappearing!

Note the subscripts being used to distinguish between the "forward" and "backward" rate constants.

Incidentally, another way to symbolize the two reactions $X + Y \xrightarrow{k_+} Z$ and $Z \xrightarrow{k_-} X + Y$ is as follows:

$$X + Y \underset{k_-}{\overset{k_+}{\rightleftharpoons}} Z \,.$$
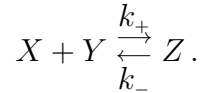
Here is one last example: $X + Y \xrightarrow{k} Z$ and $Z \xrightarrow{k'} X$ give:

$$
\begin{aligned}
dx/dt &= -kxy + k'z \\
dy/dt &= -kxy \\
dz/dt &= kxy - k'z \,.
\end{aligned}
$$

Conservation laws are often very useful in simplifying the study of chemical reactions.

For example, take the reversible bimolecular reaction that we just saw:

$$
\begin{aligned}
dx/dt &= -k_+xy + k_-z \\
dy/dt &= -k_+xy + k_-z \\
dz/dt &= k_+xy - k_-z \,.
\end{aligned}
$$

Since, clearly, $d(x + z)/dt \equiv 0$ and $d(y + z)/dt \equiv 0$, then, for every solution, there are constants x_0 and y_0 such that $x + z \equiv x_0$ and $y + z \equiv y_0$. Therefore, once that these constants are known, we only need to study the following scalar first-order ODE:

$$dz/dt = k_+(x_0 - z)(y_0 - z) - k_-z \,.$$

in order to understand the time-dependence of solutions. Once that $z(t)$ is solved for, we can find $x(t)$ by the formula $x(t) = x_0 - z(t)$ and $y(t)$ by the formula $y(t) = y_0 - z(t)$.

Note that one is only interested in non-negative values of the concentrations, which translates into the constraint that $0 \leq z \leq \min\{x_0, y_0\}$.[27]

The equation $dz/dt = k_+(x_0 - z)(y_0 - z) - k_-z$ is easily shown to have a unique, and globally asymptotically stable, positive steady state, subject to the constraint that $0 \leq z \leq \min\{x_0, y_0\}$.

(Simply intersect the line $u = k_-z$ with the parabola $u = k_+(x_0 - z)(y_0 - z)$, and use a phase-line argument: degradation is larger than production when to the right of this point, and viceversa.)

We'll see an example of the use of conservation laws when modeling enzymatic reactions.

2.6.2 Chemical Networks

We next discuss a formalism that allows one to easily write up differential equations associated with chemical reactions given by diagrams like

$$2H + O \leftrightarrow H_2O \,. \tag{2.6}$$

---

[27]This is a good place for class discussion of necessary and sufficient conditions for forward-invariance of the non-negative orthant. To be added to notes.

In general, we consider a collection of chemical reactions that involves a set of n_s "species":

$$S_i, \ i \in \{1, 2, \ldots n_s\}\,.$$

These "species" may be ions, atoms, or molecules (even large molecules, such as proteins). We'll just say "molecules", for simplicity. For example, (2.6) represents a set of two reactions that involve the following $n_s = 3$ species (hydrogen, oxygen, water):

$$S_1 = H, \quad S_2 = O, \quad S_3 = H_2O\,,$$

one going forward and one going backward. In general, a *chemical reaction network* ("CRN", for short) is a set of chemical reactions $\mathcal{R}_j, j \in \{1, 2, \ldots, n_r\}$:

$$\mathcal{R}_j : \quad \sum_{i=1}^{n_s} \alpha_{ij} S_i \to \sum_{i=1}^{n_s} \beta_{ij} S_i \tag{2.7}$$

where the α_{ij} and β_{ij} are some nonnegative integers, called the *stoichiometry coefficients*.

The species with nonzero coefficients on the left-hand side are usually referred to as the *reactants*, and the ones on the right-hand side are called the *products*, of the respective reaction. (Zero coefficients are not shown in diagrams.) The interpretation is that, in reaction 1, α_{11} molecules of species S_1 combine with α_{21} molecules of species S_2, etc., to produce β_{11} molecules of species S_1, β_{21} molecules of species S_2, etc., and similarly for each of the other $n_r - 1$ reactions.

The forward arrow means that the transformation of reactants into products only happens in the direction of the arrow. For example, the reversible reaction (2.6) is represented by the following CRN, with $n_r = 2$ reactions:

$$\mathcal{R}_1 : \quad 2H + O \to H_2O$$

$$\mathcal{R}_2 : \quad H_2O \to \quad 2H + O\,.$$
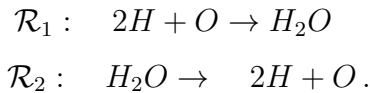
So, in this example,

$$\alpha_{11} = 2, \quad \alpha_{21} = 1, \quad \alpha_{31} = 0, \quad \beta_{11} = 0, \quad \beta_{21} = 0, \quad \beta_{31} = 1,$$

and

$$\alpha_{12} = 0, \quad \alpha_{22} = 0, \quad \alpha_{32} = 1, \quad \beta_{12} = 2, \quad \beta_{22} = 1, \quad \beta_{32} = 0\,.$$

It is convenient to arrange the stoichiometry coefficients into an $n_s \times n_r$ matrix, called the *stoichiometry matrix* $\Gamma = \Gamma_{ij}$, defined as follows:

$$\Gamma_{ij} = \beta_{ij} - \alpha_{ij}, \quad i = 1, \ldots, n_s, \quad j = 1, \ldots, n_r\,. \tag{2.8}$$

The matrix Γ has as many columns as there are reactions. Each column shows, for all species (ordered according to their index i), the net "produced$-$consumed". For example, for the reaction (2.6), Γ is the following matrix:

$$\begin{pmatrix} -2 & 2 \\ -1 & 1 \\ 1 & -1 \end{pmatrix}\,.$$

Notice that we allow degradation reactions like $A \to 0$ (all β's are zero for this reaction).

We now describe how the state of the network evolves over time, for a given CRN. We need to find a rule for the evolution of the vector:

$$\begin{pmatrix} [S_1(t)] \\ [S_2(t)] \\ \vdots \\ [S_{n_s}(t)] \end{pmatrix}$$

where the notation $[S_i(t)]$ means the concentration of the species S_i at time t. For simplicity, we drop the brackets and write S_i also for the concentration of S_i (sometimes, to avoid confusion, we use instead lower-case letters like s_i to denote concentrations). As usual with differential equations, we also drop the argument "t" if it is clear from the context. Observe that only nonnegative concentrations make physical sense (a zero concentration means that a species is not present at all).

The graphical information given by reaction diagrams is summarized by the matrix Γ. Another ingredient that we require is a formula for the actual rate at which the individual reactions take place.

We denote by $R_j(S)$ be algebraic form of the jth reaction. The most common assumption is that of *mass-action kinetics*, where:

$$R_j(S) = k_j \prod_{i=1}^{n_s} S_i^{\alpha_{ij}} \text{ for all } j = 1, \ldots, n_r .$$

This says simply that the reaction rate is proportional to the products of concentrations of the reactants, with higher exponents when more than one molecule is needed. The coefficients k_i are "reaction constants" which usually label the arrows in diagrams. Let us write the vector of reactions as $R(S)$:

$$R(S) := \begin{pmatrix} R_1(S) \\ R_2(S) \\ \vdots \\ R_{n_r}(S) \end{pmatrix}.$$

With these conventions, the system of differential equations associated to the CRN is given as follows:

$$\frac{dS}{dt} = \Gamma R(S). \tag{2.9}$$

Example

As an illustrative example, let us consider the following set of chemical reactions:

$$E + P \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} C \xrightarrow{k_2} E + Q, \qquad F + Q \underset{k_{-3}}{\overset{k_3}{\rightleftarrows}} D \xrightarrow{k_4} F + P, \tag{2.10}$$
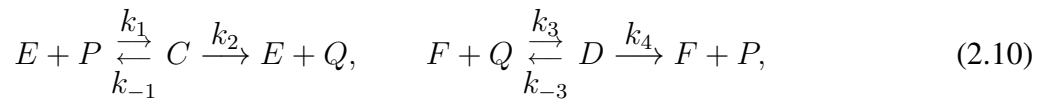
which may be thought of as a model of the activation (for instance, by phosphorylation) of a protein substrate P by an enzyme E; C is an intermediate complex, which dissociates either back into the original components or into a product (activated protein) Q and the enzyme. The second reaction transforms Q back into P, and is catalyzed by another enzyme F (for instance, a phosphatase that removes the phosphorylation). A system of reactions of this type is sometimes called a "futile cycle", and reactions of this type are ubiquitous in cell biology. The mass-action kinetics model is then

obtained as follows. Denoting concentrations with the same letters (P, etc) as the species themselves, we have the following vector of species, stoichiometry matrix Γ and vector of reaction rates $R(S)$:

$$S = \begin{pmatrix} P \\ Q \\ E \\ F \\ C \\ D \end{pmatrix}, \quad \Gamma = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & -1 & 1 & 0 \\ -1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 1 \\ 1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & -1 \end{pmatrix} \quad R(S) = \begin{pmatrix} k_1 EP \\ k_{-1} C \\ k_2 C \\ k_3 FQ \\ k_{-3} D \\ k_4 D \end{pmatrix}.$$

From here, we can write the equations (2.9). For example,

$$\frac{dP}{dt} = (-1)(k_1 EP) + (1)(k_{-1} C) + (1)(k_4 D) = k_4 D - k_1 EP + k_{-1} C.$$

Conservation Laws

Let us consider the set of row vectors c such that $c\Gamma = 0$. Any such vector is a *conservation law*, because

$$\frac{d(cS)}{dt} = c\frac{dS}{dt} = c\Gamma R(S) = 0$$

for all t, in other words,

$$c\, S(t) = \text{constant}$$

along all solutions (a "first integral" of the motion). The set of such vectors forms a linear subspace (of the vector space consisting of all row vectors of size n_s).

For instance, in the previous example, we have that, along all solutions, one has that

$$P(t) + Q(t) + C(t) + D(t) \equiv \text{constant}$$

because $(1, 1, 0, 0, 1, 1)\Gamma = 0$. Similarly, we have two more linearly independent conservation laws, namely $(0, 0, 1, 0, 1, 0)$ and $(0, 0, 0, 1, 0, 1)$, so also

$$E(t) + C(t) \quad \text{and} \quad F(t) + D(t)$$

are constant along trajectories. Since Γ has rank 3 (easy to check) and has 6 rows, its left-nullspace has dimension three. Thus, a basis of the set of conservation laws is given by the three that we have found.

2.6.3 Introduction to Enzymatic Reactions

Catalysts facilitate reactions, converting *substrates* into *products*, while remaining basically unchanged.

Catalysts may act as "pliers" that place an appropriate stress to help break a bond, they may bring substrates together, or they may help place a chemical group on a substrate.

enzyme (in substrate–binding and product–binding conformations)

substrate

product
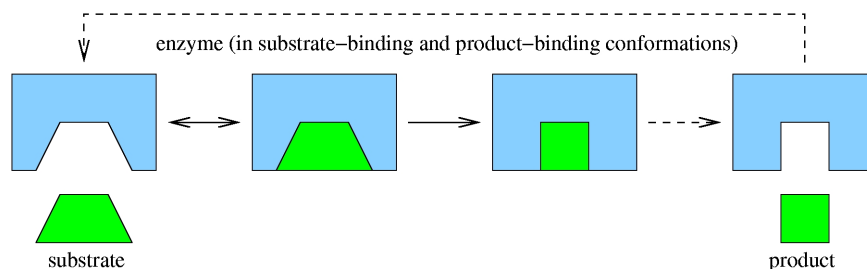
In molecular biology, certain types of proteins, called *enzymes*, act as catalysts.

Enzymatic reactions are one of the main ways in which information flows in cells.

One important type of enzymatic reaction is *phosphorylation*, when an enzyme X (called a *kinase* when playing this role) transfers a phosphate group (PO_4) from a "donor" molecule such as ATP to another protein Y, which becomes "activated" in the sense that its energy is increased.
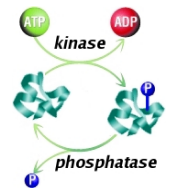
(Adenosine triphosphate (ATP) is a nucleotide that is the major energy currency of the cell:
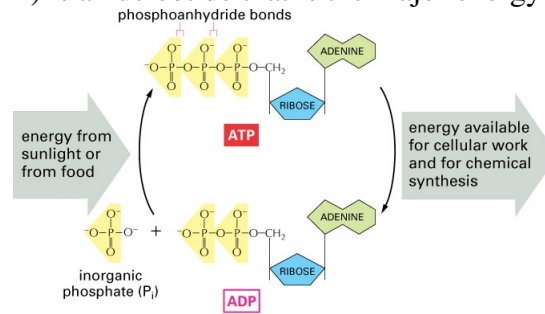
Figure from Essential Cell Biology, Second Edition, published by Garland Science in 2004; ⓒby Alberts et al

Once activated, protein Y may then influence other cellular components, including other proteins, acting itself as a kinase.

Normally, proteins do not stay activated forever; another type of enzyme, called a *phosphatase*, eventually takes away the phosphate group.

In this manner, signaling is "turned off" after a while, so that the system is ready to detect new signals.

Chemical and electrical signals from the outside of the cell are sensed by *receptors*.

Receptors are proteins that act as the cell's sensors of outside conditions, relaying information to the inside of the cell.

In some ways, receptors may be viewed as enzymes: the "substrate" is an extracellular ligand (a molecule, usually small, outside the cell, for instance a hormone or a growth factor), and the "product' might be, for example, a small molecule (a *second messenger*) that is released in response to the binding of ligand to the receptor. (Or, we may view a new conformation of the receptor as the "product" of the reaction.)

This release, in turn, may trigger signaling through a series of chemical reactions inside the cell.

Cascades and feedbacks involving enzymatic (and other) reactions, as well as the action of proteins on DNA (directing transcription of genes) are "life".

Below we show one signaling pathway, extracted from a recent paper by Hananan and Weinberg on cancer research. It describes the top-level schematics of the wiring diagram of the circuitry (in mammalian cells) responsible for growth, differentiation, and apoptosis (commands which instruct the cell to die). Highlighted in red are some of the genes known to be functionally altered in cancer cells. Almost all the main species shown are proteins, acting many of them as enzymes in catalyzing "downstream" reactions.
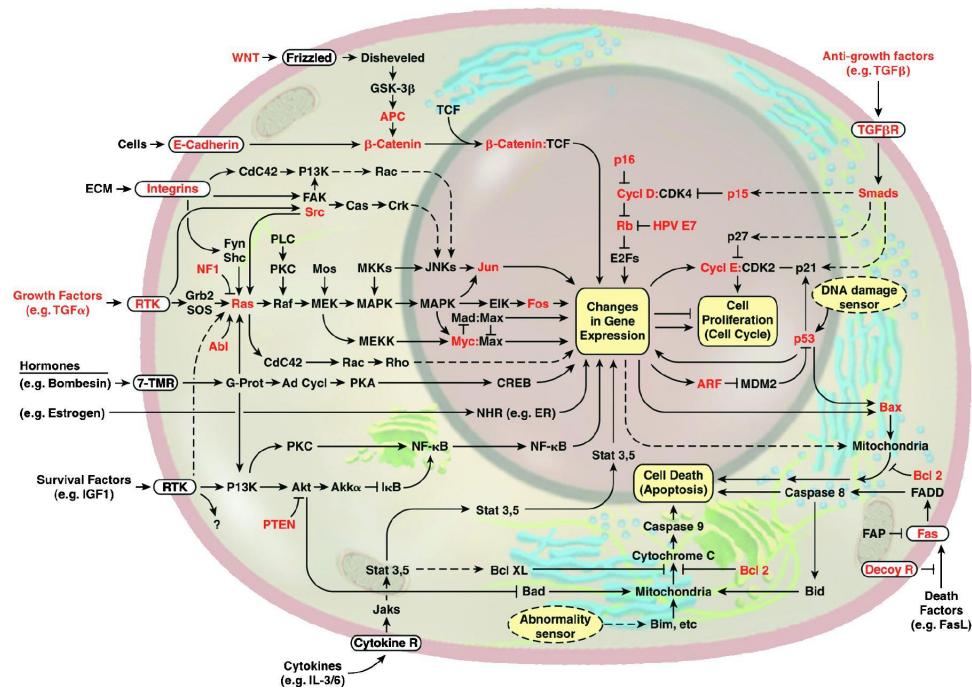
Some More on Receptors

As shown in the above diagram, most receptors are designed to recognize a specific type of ligand.

Receptors are usually made up of several parts.

• An extracellular domain ("domains" are parts of a protein) is exposed to the exterior of the cell, and this is where ligands bind.

• A transmembrane domain serves to "anchor" the receptor to the cell membrane.

• A cytoplasmic domain helps initiate reactions inside the cell in response to exterior signals, by interacting with other proteins.

As an example, a special class of receptors which constitute a common target of pharmaceutical drugs are G-protein-coupled receptors (GPCR's).

The name of these receptors arises from the fact that, when their conformation changes in response to a ligand binding event, they activate G-proteins, so called because they employ guanine triphosphate (GTP) and guanine diphosphate (GDP) in their operation.

GPCR's are made up of several subunits (G_α, G_β, G_γ) and are involved in the detection of metabolites, odorants, hormones, neurotransmitters, and even light (rhodopsin, a visual pigment).
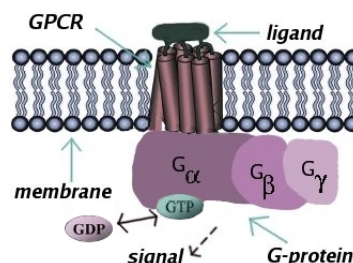
2.6.4 Differential Equations

The basic elementary reaction is:

$$S + E \; \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \; C \; \overset{k_2}{\longrightarrow} \; P + E$$
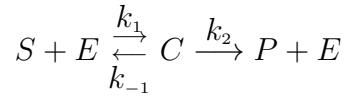
and therefore the equations that relate the concentrations of substrate, (free) enzyme, complex (enzyme with substrate together), and product are:

$$\frac{ds}{dt} = k_{-1}c - k_1 se$$

$$\frac{de}{dt} = (k_{-1} + k_2)c - k_1 se$$

$$\frac{dc}{dt} = k_1 se - (k_{-1} + k_2)c$$

$$\frac{dp}{dt} = k_2 c$$

which is a 4-dimensional system.

Since the last equation, for product formation, does not feed back into the first three,
we can simply ignore it at first, and later, after solving for $c(t)$, just integrate so as to get $p(t)$.

Moreover, since $\frac{de}{dt} + \frac{dc}{dt} \equiv 0$, we also know that $e + c$ is constant. We will write "e_0" for this sum:

$$e(t) + c(t) = e_0 .$$

(Often $c(0) = 0$ (no substrate), so that $e_0 = e(0)$, the initial concentration of free enzyme.)

So, we can eliminate e from the equations:

$$\frac{ds}{dt} = k_{-1}c - k_1 s(e_0 - c)$$

$$\frac{dc}{dt} = k_1 s(e_0 - c) - (k_{-1} + k_2)c .$$

We are down to two dimensions, and could proceed using the methods that we have been discussing.

However, Leonor Michaelis and Maud Leonora Menten formulated in 1913 an approach that allows one to reduce the problem even further, by doing an approximation. Next, we review this approach, as reformulated by Briggs and Haldane in 1925[28], and interpret it in the more modern language of singular perturbation theory.

Although a two-dimensional system is not hard to study, the reduction to one dimension is very useful:
• When "connecting" many enzymatic reactions, one can make a similar reduction for each one of the reactions, which provides a great overall reduction in complexity.
• It is often not possible, or it is very hard, to measure the kinetic constants (k_1, etc), but it may be easier to measure the parameters in the reduced model.

---

[28]Michaelis and Menten originally made an the "equilibrium approximation" $k_{-1}c(t) - k_1 s(t)e(t) = 0$ in which one assumes that the first reaction is in equilibrium. This approximation is very hard to justify. The Briggs and Haldane approach makes a different approximation. The final form of the production rate (see later) turns out to be algebraically the same as in the original Michaelis and Menten work, but the parameters have different physical interpretations in terms of the elementary reactions.

2.6.5   Quasi-Steady State Approximations and Michaelis-Menten Reactions

Let us write

$$
\frac{ds}{dt} = k_{-1}c - k_1 s(e_0 - c)
$$

$$
\frac{dc}{dt} = k_1 s(e_0 - c) - (k_{-1} + k_2)c = k_1\left[s\,e_0 - (K_{\mathrm{m}} + s)c\right], \qquad \text{where} \quad K_{\mathrm{m}} = \frac{k_{-1} + k_2}{k_1}.
$$

The MM approximation amounts to setting $dc/dt = 0$. The biochemical justification is that, after a transient period during which the free enzymes "fill up," the amount complexed stays more or less constant.

This allows us, by solving the algebraic equation:

$$
s\,e_0 - (K_{\mathrm{m}} + s)c = 0
$$

to express c in terms of s:

$$
c = \frac{s\,e_0}{K_{\mathrm{m}} + s}. \tag{2.11}
$$

We then have, for the production rate:

$$
\frac{dp}{dt} = k_2\,c = \frac{V_{\max}s}{K_{\mathrm{m}} + s}. \tag{2.12}
$$

Also, substituting into the s equation we have:

$$
\frac{ds}{dt} = k_{-1}\frac{s\,e_0}{K_{\mathrm{m}} + s} - k_1 s\left(e_0 - \frac{s\,e_0}{K_{\mathrm{m}} + s}\right) = -\frac{V_{\max}s}{K_{\mathrm{m}} + s} \tag{2.13}
$$

where we denote $V_{\max} = k_2 e_0$.   If we prefer to explicitly show the role of the enzyme as an "input", we can write these two equations as follows:

$$
\frac{ds}{dt} = -e_0\frac{k_2\,s}{K_{\mathrm{m}} + s}
$$

$$
\frac{dp}{dt} = e_0\frac{k_2\,s}{K_{\mathrm{m}} + s}
$$

showing the rate at which substrate gets transformed into product with the help of the enzyme.

This is all very nice, and works out well in practice, but the mathematical justification is flaky: setting $dc/dt = 0$ means that c is constant. But then, the equation $c = \frac{s\,e_0}{K_{\mathrm{m}}+s}$ implies that s must be constant, too. Therefore, also $ds/dt = 0$.

But then $\frac{V_{\max}s}{K_{\mathrm{m}}+s} = -ds/dt = 0$, which means that $s = 0$. In other words, our derivation can only be right if there is no substrate, so no reaction is taking place at all!

One way to justify these derivations is as follows. Under appropriate conditions, s changes much more slowly than c.

So, as far as c is concerned, we may assume that $s(t)$ is constant, let us say $s(t) = \bar{s}$.

Then, the equation for c becomes a linear equation, which converges to its steady state, which is given by formula (2.11) (with $s = \bar{s}$) obtained by setting $dc/dt = 0$.

Now, as s changes, c "catches up" very fast, so that this formula is always (approximately) valid.

From the "point of view" of s, the variable c is always catching up with its expression given by formula (2.11), so, as far as its slow movement is concerned, s evolves according to formula (2.13). (An exception is at the start of the whole process, when $c(0)$ is initially far from its steady state value. This is the "boundary layer behavior".)

2.6.6 A quick intuition with nullclines

Let us introduce the following rescaled variables:

$$x = \frac{s}{s_0}, \quad y = \frac{c}{e_0},$$

and write also $\varepsilon = e_0/s_0$, where we think of s_0 as the initial concentration $s(0)$ of substrate. *We will make the assumption that the initial concentration e_0 of enzyme e is small compared to that of substrate, i.e. that the ratio ε is small*[29]. Note that x, y, ε are "non-dimensional" variables.

It is clear from the equations that, if we start with $c(0) = 0$, then $s(t)$ is always $\leq s_0$, and $c(t)$ is always $\leq e_0$. Therefore, $0 \leq x(t) \leq 1$ and $0 \leq y(t) \leq 1$.

Using these new variables, the equations become:

$$\frac{dx}{dt} = \varepsilon \left[k_{-1} y - k_1 s_0 x (1 - y) \right]$$
$$\frac{dy}{dt} = k_1 \left[s_0 x - (K_m + s_0 x) y \right].$$

The y nullcline is the graph of:

$$y = \frac{s_0 x}{K_m + s_0 x} \tag{2.14}$$

(which is the same as saying that the c nullcline is the graph of $c = \dfrac{s e_0}{K_m + s}$) and the x nullcline is the graph of:

$$y = \frac{s_0 x}{\frac{k_{-1}}{k_1} + s_0 x}. \tag{2.15}$$

Now, since $\frac{k_{-1}}{k_1} < \frac{k_{-1}+k_2}{k_1} = K_m$, it follows that the y-nullcline lies under the x-nullcline.

In addition, *using that ε is small*, we can say that the vector field should be quite "vertical" (small x component compared to y component), at least if we are far from the y-nullcline.

(The x component is small because ε is small, since x and y are both bounded by one. The y component will be ≈ 0 when we are near the y-nullcline.)

It is easy to see then that the phase plane looks as follows, qualitatively (two typical trajectories are shown):

---

[29]It would not make sense to just say that the amount of enzyme is "small," since the meaning of "small" depends on units. On the other hand, the *ratio* makes sense, assuming of course that we quantified concentrations of enzyme and substrate in the same units. Typical values for ε may be in the range 10^{-2} to 10^{-7}.
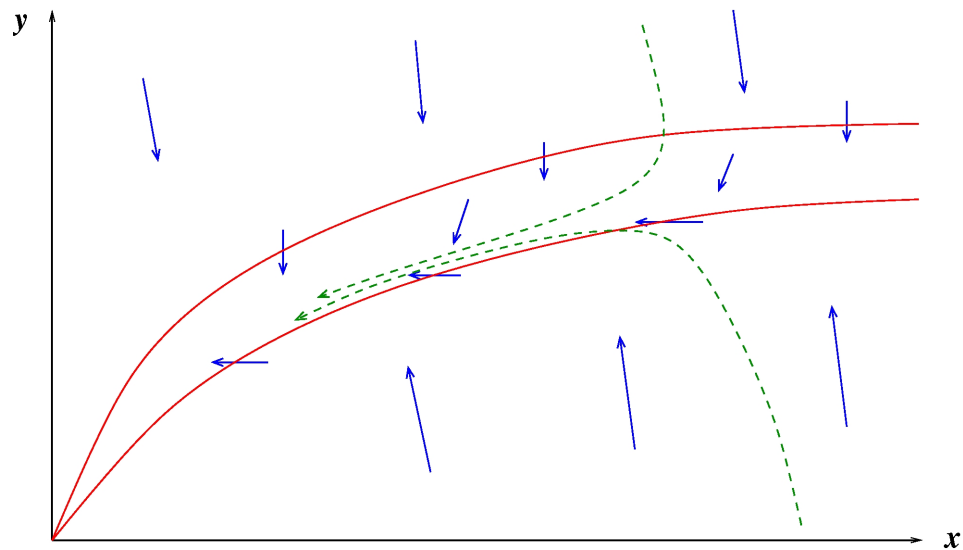
In fact, with $s_0 = e_0 = k_- = k_{-1} = k_2 = 1$, this is the actual phase plane (nullclines and vector field shown).
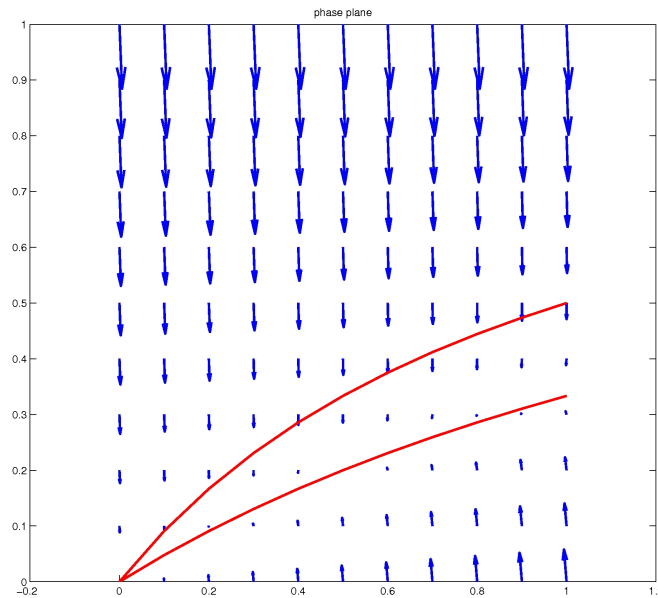
It was generated using the following MATLAB code:

```
eps = 0.1;s0 = 1;
[X,Y]=meshgrid(0:0.1:1, 0:0.1:1);
z1= eps.*(Y - s0.*X.*(1-Y));z2= s0.*X-((2+s0.*X).*Y);
quiver(X,Y,z1,z2,'LineWidth',2)
title('phase plane')
hold;
x=0:0.1:1;
plot(x,x./(2.+x),'LineWidth',2,'color','r')
plot(x,x./(1.+x),'LineWidth',2,'color','r')
```

The key point is that (in these coordinates) trajectories initially move almost "vertically" toward the y-nullcline, and subsequently stay very close to this nullcline, for large times t. This means that, for large t, $c(t) \approx \dfrac{s(t)\, e_0}{K_m + s(t)}$, which is what the MM approximation (2.11) claims.

Using "$c(t) = \frac{s(t)\,e_0}{K_m + s(t)}$" is called a "*quasi*-steady state approximation," because it formally looks like "$dc/dt = 0$," which would be like saying that the c component is at the value that it would be if the system were at steady state (which it really isn't).

To make all this more precise mathematically, one needs to do a "time scale analysis" which studies the dynamics from c's point of view (slow time scale) and s's (fast time scale) separately. The next few sections provide some more details. The reader may wish to skip to subsection 2.6.9.

2.6.7   Fast and Slow Behavior

Let us start again from the equations

$$\frac{dx}{dt} = \varepsilon\left[k_{-1}\,y - k_1 s_0\,x\,(1 - y)\right]$$

$$\frac{dy}{dt} = k_1\left[s_0\,x - (K_m + s_0\,x)y\right].$$

in the coordinates $x = \frac{s}{s_0},\ y = \frac{c}{e_0}$.

Since $\varepsilon \approx 0$, we make the approximation "$\varepsilon = 0$" and substitute $\varepsilon = 0$ into these equations. (Note that x and y are bounded by 1, so they remain bounded.)

So $dx/dt = 0$, which means that $x(t)$ equals a constant $\bar{x}$, and hence the second equation becomes:

$$\frac{dc}{dt} = k_1\left[e_0\,\bar{s} - (K_m + \bar{s})c\right]$$

(substituting $s_0 x = s$ and $e_0 y = c$ to express in terms of the original variables, and letting $\bar{s} = s_0\bar{x}$). In this differential equation, $c(t)$ converges as $t \to \infty$ to the steady state

$$c = \frac{e_0\,\bar{s}}{K_m + \bar{s}}$$

which is also obtained by setting $dc/dt = 0$ in the original equations if $s(t) \equiv \bar{s}$ is assumed constant.

(Observe that the speed of convergence is determined by $k_1(K_m + \bar{s})$, which does not get small as $\varepsilon \to 0$.)

In this way, we again obtain formula (2.12) for dp/dt ($\bar{s}$ is the "present" value of s).

This procedure is called a "quasi-steady state approximation" (QSS), reflecting the fact that one replaces c by its "steady state" value $\frac{e_0\,s}{K_m + s}$ obtained by pretending that s would be constant. This is not a true steady state of the original equations, of course.

In summary, assuming $\varepsilon \approx 0$, we made the approximation "$\varepsilon = 0$" leading to $dx/dt \equiv 0$ and $x(t) \equiv \bar{x}$.

However, "$\varepsilon \approx 0$" is not the same as "$\varepsilon = 0$", so we cannot really say that $dx/dt = 0$. Eventually, $x(t)$ changes!

Yet, the idea still works, but we need to make a more careful argument using *time-scale separation*: the key point is that c approaches its steady state value fast relative to the movement of s, which may, therefore, supposed to be constant while this convergence happens.

So we "iterate" the reasoning: s moves a bit, using c's steady state value. Then, c "reacts" to this new value of s, converging to a new steady state value (corresponding to the new $\bar{s}$), and the process is iterated in this fashion.

The main problem with saying things in this manner is that, of course, it is not true that c and s take turns moving, but both move simultaneously (although at very different speeds).

Long-time behavior (fast time scale)

In order to be more precise, it is convenient to make a change of time scale, using:

$$\tau = \frac{e_0}{s_0} k_1 t \,.$$

We may think of τ as a *fast time scale*, because $\tau = \varepsilon k_1 t$, and therefore τ is small for any given t.

For example, if $\varepsilon k_1 = 1/3600$ and t is measured in seconds, then $\tau = 10$ implies that $t = 36000$; thus, "$\tau = 10$" means that ten *hours* have elapsed, while "$t = 10$" means that only ten seconds elapsed.

Substituting $s = s_0 x$, $c = e_0 y$, and

$$\frac{dx}{d\tau} = \frac{1}{e_0 k_1} \frac{ds}{dt} \,, \quad \frac{dy}{d\tau} = \frac{s_0}{e_0^2 k_1} \frac{dc}{dt} \,,$$

we have:

$$\frac{dx}{d\tau} = \frac{k_{-1}}{k_1} y - s_0 x (1 - y)$$

$$\varepsilon \frac{dy}{d\tau} = s_0 x - (K_{\mathrm{m}} + s_0 x) y \,.$$

Still assuming that $\varepsilon \ll 1$, we make an approximation by setting $\varepsilon = 0$ in the second equation:

$$\varepsilon \frac{dy}{d\tau} = s_0 x - (K_{\mathrm{m}} + s_0 x) y$$

leading to the algebraic equation $s_0 x - (K_{\mathrm{m}} + s_0 x) y = 0$ which we solve for $y = y(x) = \frac{s_0 x}{K_{\mathrm{m}} + s_0 x}$, or equivalently

$$c = \frac{e_0 s}{K_{\mathrm{m}} + s} \,, \tag{2.16}$$

and finally we substitute into the first equation:

$$\frac{dx}{d\tau} = \frac{k_{-1}}{k_1} y - s_0 x (1 - y) = -\frac{(-k_{-1} + K_{\mathrm{m}} k_1) s_0 x}{k_1 (K_{\mathrm{m}} + s_0 x)} = -\frac{k_2 s_0 x}{k_1 (K_{\mathrm{m}} + s_0 x)}$$

(recall that $K_{\mathrm{m}} = \frac{k_{-1} + k_2}{k_1}$).

In terms of the original variable $s = s_0 x$, using $\frac{ds}{dt} = e_0 k_1 \frac{dx}{d\tau}$, and recalling that $V_{\max} = k_2 e_0$, we have re-derived (2.13):

$$\frac{ds}{dt} = -\frac{V_{\max} s}{K_{\mathrm{m}} + s} \,.$$

The important point to realize is that, after an initial convergence of c (or y) to its steady state, once that c has "locked into" its steady state (2.16), it quickly "catches up" with any (slow!) changes in s, and this catch-up is not "visible" at the time scale τ, so c appears to track the expression (2.16).

Short initial-time behavior (slow time scale)

One special case is that of small initial times t, when c (or y) has not yet converged to a steady state. For $t \approx 0$, we may assume that $\bar{s} = s_0$, and therefore the equation for c is approximated by:

$$\frac{dc}{dt} = k_1 \left[e_0 \, s_0 - (K_{\mathrm{m}} + s_0)c \right] . \tag{2.17}$$

One calls this the *boundary layer* equation, because it describes what happens near initial times (boundary of the time interval).

Putting it all Together

Let's suppose that $s(0) = s_0$ and $c(0) = c_0$.

(1) As we remarked earlier, for $t \approx 0$ we have equation (2.17) (with initial condition $c(0) = c_0$).

(2) For t large , we have the approximations given by (2.16) for c, and (2.13) for s.

The "Method of Matched Asymptotic Expansions" (not covered here) is used to patch the inner or boundary-layer solution with the outer or fast time scale solution in order to obtain a globally valid solution.

The approximation is best if ε is very small, but it works quite well even for moderate ε.
Here is a numerical example.

Let us take $k_1 = k_{-1} = k_2 = e_0 = 1$ and $s_0 = 10$, so that $\varepsilon = 0.1$. Note that $K_{\mathrm{m}} = 2$ and $V_{\max} = 1$.

We show below, together, the following plots:

• in black, the component $c(t)$ of the true solution of the system

$$\frac{ds}{dt} = c - s(1 - c), \quad \frac{dc}{dt} = s - (2 + s)c$$

with initial conditions $s(0) = s_0$, $c(0) = 0$,

• in red, $c = s/(2 + s)$, where $s(t)$ solves $\frac{ds}{dt} = -s/(2 + s)$ (slow system) with $s(0) = s_0$,

• in blue, the solution of the fast system at the initial time, $\frac{dc}{dt} = s_0 - (2 + s_0)c$, with $c(0) = 0$.

Since it is difficult to see the curves for small t, we show plots both for $t \in [0, 25]$ and for $t \in [0, 0.5]$:
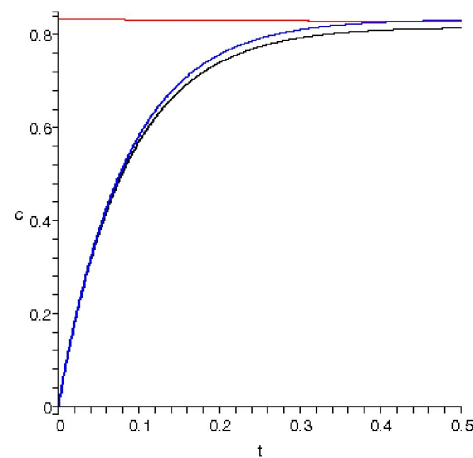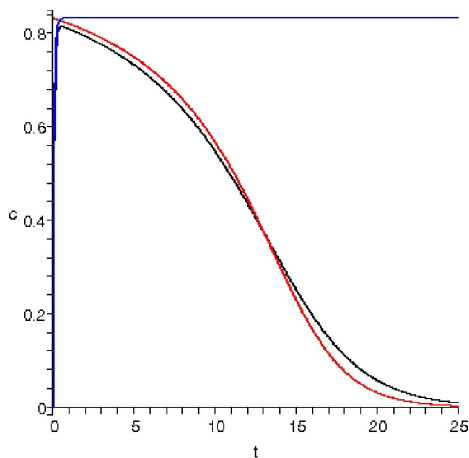
As expected, the blue curve approximates well for small t and the red one for larger t.

FYI, here is the Maple code that was used (for Tmax = 0.5 and 25):

```
restart:with(plots):with(DEtools):
s0:=10:Tmax:=0.5:N:=500:
sys:=diff(s(t),t)=c(t)-s(t)*(1-c(t)),diff(c(t),t)=s(t)-(2+s(t))*c(t):
sol:=dsolve(sys,s(0)=s0,c(0)=0,type=numeric):
plot1:=odeplot(sol,[[t,c(t)]],0..Tmax,numpoints=N,color=black,thickness=3):
sysslow:= diff(s(t),t) = - s(t)/(2+s(t)):
solslow:=dsolve(sysslow,s(0)=s0,type=numeric):
solns:= t → op(2,op(2,solslow(t))):
plot2:=plot(solns/(2+solns),0..Tmax,numpoints=N,color=red,thickness=3):
sysfast:=diff(c(t),t)=s0-(2+s0)*c(t):
solfast:=dsolve(sysfast,c(0)=0,type=numeric):
plot3:=odeplot(solfast,[[t,c(t)]],0..Tmax,numpoints=N,color=blue,thickness=3):
display(plot1,plot2,plot3);
```

2.6.8 Singular Perturbation Analysis

The advantage of deriving things in this careful fashion is that we have a better understanding of what went into the approximations. Even more importantly, there are methods in mathematics that help to quantify the errors made in the approximation. The area of mathematics that deals with this type of argument is *singular perturbation theory*.

The theory applies, in general, to equations like this:

$$\frac{dx}{dt} = f(x, y)$$
$$\varepsilon \frac{dy}{dt} = g(x, y)$$

with $0 < \varepsilon \ll 1$. The components of the vector x are called *slow* variables, and those of y *fast* variables.

The terminology is easy to understand: $dy/dt = (1/\varepsilon)(\ldots)$ means that dy/dt is large, i.e., that $y(t)$ is "fast," and by comparison $x(t)$ is slow.[30]

The singular perturbation approach starts by setting $\varepsilon = 0$,
then solving (if possible) $g(x, y) = 0$ for $y = h(x)$ (that is, $g(x, h(x)) = 0$),
and then substituting back into the first equation.

Thus, one studies the *reduced system*:

$$\frac{dx}{dt} = f(x, h(x))$$

on the "slow manifold" defined by $g(x, y) = 0$.

---

[30]The theory covers also multiple, not just two, time scales, as well partial differential equations where the domain is subject to small deformations, and many other situations as well.

Slow manifold g(x,y)=0. If starting on this manifold, the equations are dx/dt = f(x,h(x)) and y = h(x)

true trajectory

fast motion approximation dy/dt = g(x, y_0)

x_0
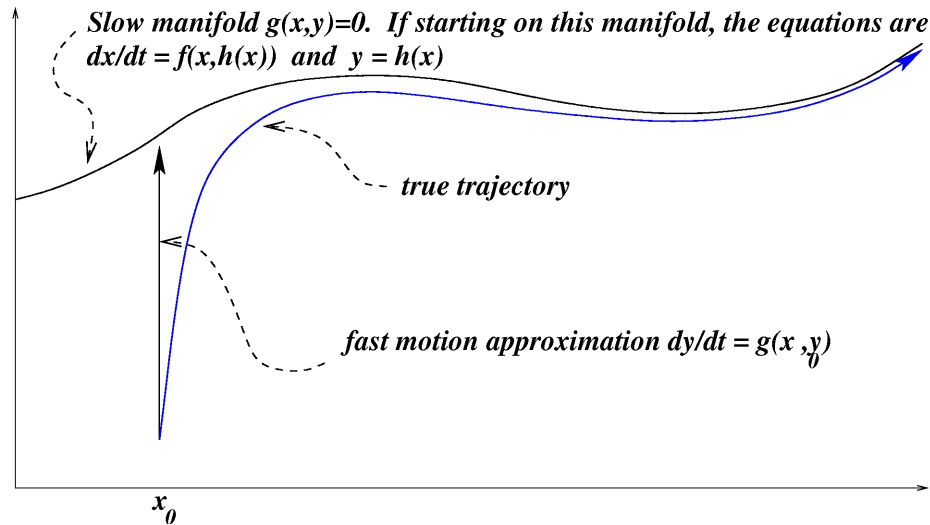
There is a rich theory that allows one to mathematically justify the approximations.

A particularly useful point of view us that of "geometric singular perturbation theory." We will not cover any of that in this course, though.

2.6.9   Inhibition

Let us discuss next inhibition, as a further example involving enzymes.

In *competitive inhibition*, a second substrate, called an inhibitor, is capable of binding to an enzyme, thus block binding of the primary substrate.

enzyme

competitive
inhibitor bound
to enzyme
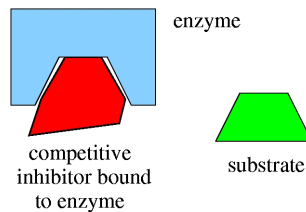
substrate

If the primary substrate cannot bind, no "product" (such as the release of signaling molecules by a receptor) can be created.

For example, the enzyme may be a cell surface receptor, and the primary substrate might be a growth factor, hormone, or histamine (a protein released by the immune system in response to pollen, dust, etc).

Competitive inhibition is one mechanism by which drugs act. For example, an inhibitor drug will attempt to block the binding of the substrate to receptors in cells that can react to that substrate, such as for example histamines to lung cells. Many antihistamines work in this fashion, e.g. Allegra.[31]

A simple chemical model is as follows:

$$S + E \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} C_1 \overset{k_2}{\longrightarrow} P + E \qquad I + E \underset{k_{-3}}{\overset{k_3}{\rightleftarrows}} C_2$$

---
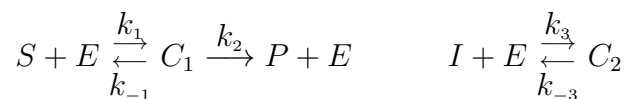
[31]In pharmacology, an *agonist* is a ligand which, when bound to a receptor, triggers a cellular response. An *antagonist* is a competitive inhibitor of an agonist. when we view the receptor as an enzyme and the agonist as a substrate.

where C_1 is the substrate/enzyme complex, C_2 the inhibitor/enzyme complex, and I the inhibitor.
In terms of ODE's, we have:

$$\frac{ds}{dt} = k_{-1}c_1 - k_1 se$$

$$\frac{de}{dt} = (k_{-1} + k_2)c_1 + k_{-3}c_2 - k_1 se - k_3 ie$$

$$\frac{di}{dt} = k_{-3}c_2 - k_3 ie$$

$$\frac{dc_1}{dt} = k_1 se - (k_{-1} + k_2)c_1$$

$$\frac{dc_2}{dt} = k_3 ie - k_{-3}c_2$$

$$\frac{dp}{dt} = k_2 c_1 \,.$$

It is easy to see that $c_1 + c_2 + e$ is constant (it represents the total amount of free or bound enzyme, which we'll denote as e_0). This allows us to eliminate e from the equations. Furthermore, as before, we may first ignore the equation for p. We are left with a set of four ODE's:

$$\frac{ds}{dt} = k_{-1}c_1 - k_1 s(e_0 - c_1 - c_2)$$

$$\frac{di}{dt} = k_{-3}c_2 - k_3 ie$$

$$\frac{dc_1}{dt} = k_1 s(e_0 - c_1 - c_2) - (k_{-1} + k_2)c_1$$

$$\frac{dc_2}{dt} = k_3 i(e_0 - c_1 - c_2) - k_{-3}c_2 \,.$$

(We could also use a conservation law $i + c_2 \equiv i_0 =$ total amount of inhibitor, free or bound to enzyme, to reduce to just three equations, but it is better for time-scale separation purposes not to do so.) One may now do a quasi-steady-state approximation, assuming that the enzyme concentrations are small relative to substrate, amounting formally to setting $dc_1/dt = 0$ and $dc_2/dt = 0$. Doing so gives:

$$c_1 = \frac{K_{\mathbf{i}} e_0 s}{K_{\mathrm{m}} i + K_{\mathbf{i}} s + K_{\mathrm{m}} K_{\mathbf{i}}} \qquad \left(K_{\mathrm{m}} = \frac{k_{-1} + k_2}{k_1} \right)$$

$$c_2 = \frac{K_{\mathrm{m}} e_0 i}{K_{\mathrm{m}} i + K_{\mathbf{i}} s + K_{\mathrm{m}} K_{\mathbf{i}}} \qquad \left(K_{\mathbf{i}} = \frac{k_{-3}}{k_3} \right) \,.$$

The product formation rate is $dp/dt = k_2 c_1$, so, again with $V_{\mathrm{max}} = k_2 e_0$, one has the approximate formula:

$$\frac{dp}{dt} = \frac{V_{\mathrm{max}} \, s}{s + K_{\mathrm{m}}(1 + i/K_{\mathbf{i}})}$$

The formula reduces to the previous one if there is no inhibitor ($i = 0$).

We see that the rate of product formation is smaller than if there had been no inhibition, given the same amount of substrate $s(t)$ (at least if $i \gg 1$, $k_3 \gg 1$, $k_{-3} \ll 1$).

But for s very large, the rate saturates at $dp/dt = V_{\mathrm{max}}$, just as if there was no inhibitor (intuitively, there is so much s that i doesn't get chance to bind and block). *Thus, to affect the amount of product being formed when the substrate amounts are large, potentially a huge amount of drug (inhibitor) would have to be administered!* Allosteric inhibition, described next, does not have the same disadvantage.

2.6.10   Allosteric Inhibition

In *allosteric inhibition*[32], an inhibitor does not bind in the same place where the catalytic activity occurs, but instead binds at a different *effector* site (other names are *regulatory* or *allosteric* site), with the result that the shape of the enzyme is modified. In the new shape, it is harder for the enzyme to bind to the substrate.
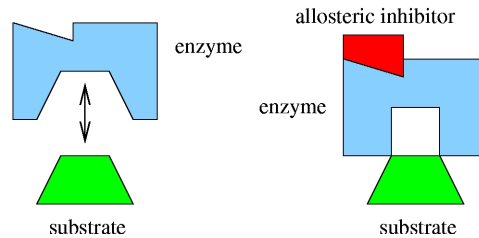
A slightly different situation is if binding of substrate can always occur, but product can only be formed (and released) if I is not bound. We model this last situation, which is a little simpler.
Also, for simplicity, we will assume that binding of S or I to E are independent of each other.
(If we don't assume this, the equations are still the same, but we need to introduce some more kinetic constants k's.)

A reasonable chemical model is, then:

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

$$EI + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} EIS$$

$$E + I \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} EI$$

$$ES + I \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} EIS$$
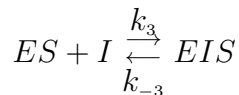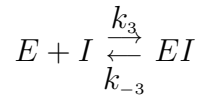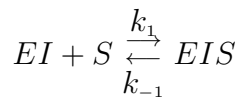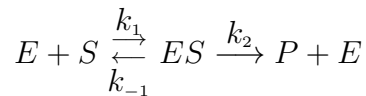
where "EI" denotes the complex of enzyme and inhibitor, etc.

It is possible to show that there results under quasi-steady state approximation a rate

$$\frac{dp}{dt} = \frac{V_{\max}}{1 + i/K_{\mathbf{i}}} \cdot \frac{s^2 + as + b}{s^2 + cx + d}$$

for some suitable numbers $a = a(i), \ldots$ and a suitably defined $K_{\mathbf{i}}$.

Notice that the maximal possible rate, for large s, is lower than in the case of competitive inhibition.

One intuition is that, no matter what is the amount of substrate, the inhibitor can still bind, so maximal throughput is affected.

---

[32]Merriam-Webster: allosteric: "all+steric"; and steric means "relating to or involving the arrangement of atoms in space" and originates with the word "solid" in Greek

2.6.11 A digression on gene expression

A very simple model of gene expression is as follows.

We let D, M, and P denote respectively the concentration of active promoter sites ("concentration" in the sense of proportion of active sites in a population of cells), mRNA transcript, and protein.

The network of reactions is:

$$D \xrightarrow{\alpha} D + M \,, \quad M \xrightarrow{\beta} 0 \,, \quad M \xrightarrow{\theta} M + P \,, \quad P \xrightarrow{\delta} 0$$
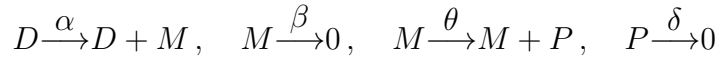
which represent, respectively, transcription and degradation of mRNA, translation, and degradation (or dilution due to cell growth) in protein concentrations.

Remark: This model ignores a huge amount of biochemistry and biophysics, such as the dynamics of mRNA polymerase's transcriptional process.
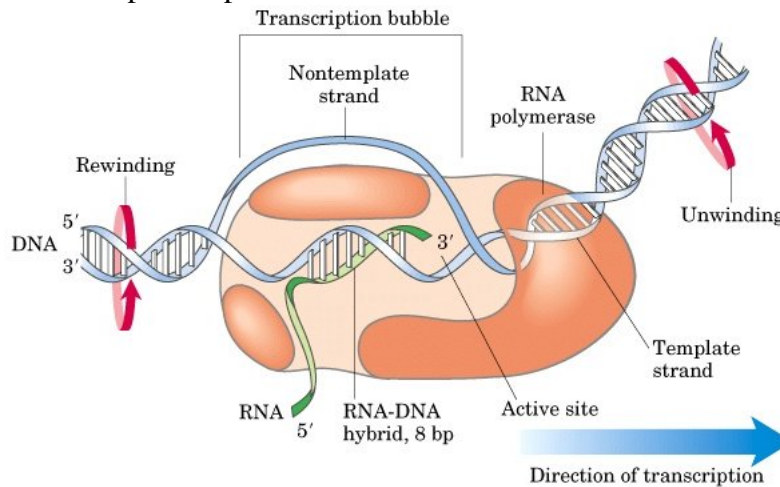
Nonetheless, it is a very useful model, and the one most often employed.

Using mass-action kinetics, we have the following rates:

$$R_1 = \alpha D \,, \quad R_2 = \beta M \,, \quad R_3 = \theta M \,, \quad R_4 = \delta P$$

for some positive constants $\alpha, \beta, \theta, \delta$. The stoichiometry matrix is:

$$\Gamma = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \,.$$

Note that, since D is not being changed, we could equally well, in this model, replace the first two reactions by $0 \xrightarrow{\alpha} M \xrightarrow{\beta} 0$, and drop D from the description. However, we include D because we will consider repression below.

A *promoter region* is a part of the DNA sequence of a chromosome that is recognized by RNA polymerase. In prokaryotes, the promoter region consists of two short sequences placed respectively 35 and 10 nucleotides before the start of the gene. Eukaryotes require a far more sophisticated transcriptional control mechanism, because different genes may be only active in particular cells or tissues at particular times in an organism's life; promoters act in concert with enhancers, silencers, and other regulatory elements
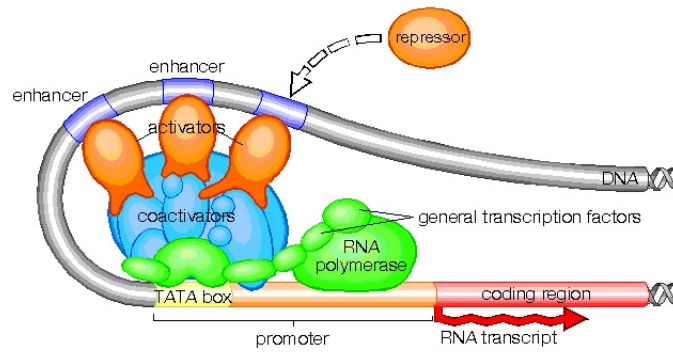
Now let's add repression to the chemical network model.

Suppose that a molecule (transcription factor) R can repress transcription by binding to DNA, hence affecting the activity of the promoter.

The model then will add an equation:

$$D + R \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} C$$
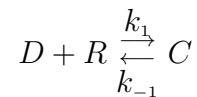
representing complex formation between promoter and repressor.

This is closely analogous to enzyme inhibition. There is an exercise that asks for an analysis of this model.

2.6.12   Cooperativity

Let's take a situation where n molecules of substrate must first get together with the enzyme in order for the reaction to take place:

$$nS + E \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} C \overset{k_2}{\longrightarrow} P + E$$
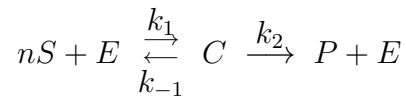
This is not a very realistic model, since it is unlikely that $n+1$ molecules may "meet" simultaneously.

It is, nonetheless, a simplification of a more realistic model in which the bindings may occur in sequence.

One says that the cooperativity degree of the reaction is n, because n molecules of S must be present for the reaction to take place.

Highly cooperative reactions are extremely common in biology, for instance, in ligand binding to cell surface receptors, or in binding of transcription factors to DNA to control gene expression.

We only look at this simple model in this course. We have these equations:

$$
\begin{aligned}
\frac{ds}{dt} &= nk_{-1}c - nk_1 s^n e \\
\frac{de}{dt} &= (k_{-1} + k_2)c - k_1 s^n e \\
\frac{dc}{dt} &= k_1 s^n e - (k_{-1} + k_2)c \\
\frac{dp}{dt} &= k_2 c
\end{aligned}
$$

Doing a quasi-steady state approximation, under the assumption that enzyme concentration is small compared to substrate, we may repeat the previous arguments and look at the c-nullcline, which leads to the same expression as earlier for product formation, except that a different exponent appears:

$$\frac{dp}{dt} = \frac{V_{\max} s^n}{K_{\mathrm{m}} + s^n}$$

The integer n is called the *Hill coefficient*.

One may determine $V_{\max}$, n, and K_{m} experimentally, from knowledge of the rate of product formation dp/dt as a function of current substrate concentration (under the quasi-steady state approximation assumption).

First, $V_{\max}$ may be estimated from the rate dp/dt corresponding to $s \to \infty$. This allows the computation of the quantity $\frac{dp/dt}{V_{\max}-dp/dt}$. Then, one observes that the following equality holds (solve for s^n and take logs):

$$n \ln s = \ln K_{\mathrm{m}} + \ln \left(\frac{dp/dt}{V_{\max} - dp/dt} \right).$$

Thus, by a linear regression of $\ln \left(\frac{dp/dt}{V_{\max}-dp/dt} \right)$ versus $\ln s$, and looking at slope and intersects, n and K_{m} can be estimated.

Since the cooperative mechanism may include many unknown and complicated reactions, including very complicated allosteric effects, it is not uncommon for fractional powers to be appear (even if the above model makes no sense in a fractional situation) when fitting parameters.

One often writes the product formation rate, redefining the constant K_{m}, as $\frac{dp}{dt} = \frac{V_{\max} s^n}{K_{\mathrm{m}}^n + s^n}$.

This has the advantage that, just as earlier, K_{m} has an interpretation as the value of substrate s for which the rate of formation of product is half of $V_{\max}$.

For our subsequent studies, the main fact that we observe is that, for $n > 1$, one obtains a "sigmoidal" shape for the formation rate, instead of a "hyperbolic" shape.

This is because, if $f(s) = \frac{V_{\max} s^n}{K_{\mathrm{m}}^n + s^n}$, then $f'(0) > 0$ when $n = 1$, but $f'(0) = 0$ if $n > 1$.

In other words, for $n > 1$, and as the function is clearly increasing, the graph must start with concavity-up. But, since the function is bounded, the concavity must change to negative at some point.

Here are graphs of two formation rates, one with $n = 1$ (hyperbolic) and one with $n = 3$ (sigmoidal):
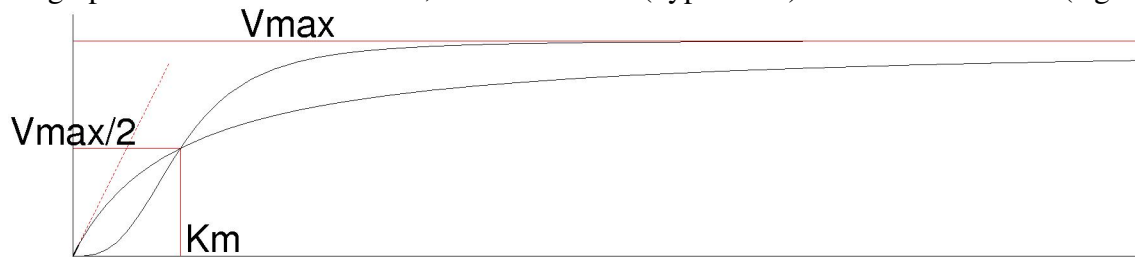
Cooperativity plays a central role in allowing for multi-stable systems, memory, and development, as we'll see soon.

Here is a more or less random example from the literature[33] which shows fits of $V_{\max}$ and n ("n_H" for "Hill") to various data sets corresponding to an allosteric reaction.

---

[33] Ian J. MacRae et al., "Induction of positive cooperativity by amino acid replacements within the C-terminal domain of Penicillium chrysogenum ATP sulfurylase," J. Biol. Chem., Vol. 275, 36303-36310, 2000

(Since you asked: the paper has to do with an intracellular reaction having to do with the incorporation of inorganic sulfate into organic molecules by sulfate assimilating organisms; the allosteric effector is PAPS, 3'-phosphoadenosine-5'-phosphosulfate.)
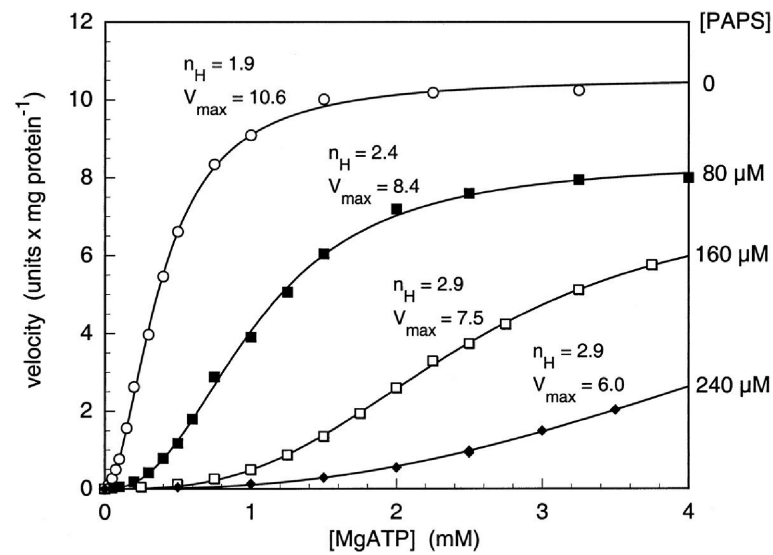
The fit to the Hill model is quite striking.

2.7 Multi-Stability

2.7.1 Hyperbolic and Sigmoidal Responses

Let us now look at the enzyme model again, but this time assuming that the substrate is not being depleted.

This is not as strange a notion as it may seem.

For example, in receptor models, the "substrate" is ligand, and the "product" is a different chemical (such as a second messenger released inside the cell when binding occurs), so the substrate is not really "consumed."

Or, substrate may be replenished and kept at a certain level by another mechanism.

Or, the change in substrate may be so slow that we may assume that its concentration remains constant.

In this case, instead of writing

$$S + E \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} C \xrightarrow{k_2} P + E \, ,$$
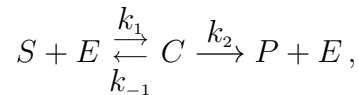
it makes more sense to write

$$E \underset{k_{-1}}{\overset{k_1 s}{\rightleftharpoons}} C \xrightarrow{k_2} P + E \, .$$
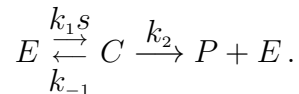
The equations are as before:

$$
\begin{aligned}
\frac{de}{dt} &= (k_{-1} + k_2)c - k_1 se \\
\frac{dc}{dt} &= k_1 se - (k_{-1} + k_2)c \\
\frac{dp}{dt} &= k_2 c
\end{aligned}
$$

except for the fact that we view s as a constant.

Repeating exactly all the previous steps, a quasi-steady state approximation leads us to the product formation rate:

$$\frac{dp}{dt} = \frac{V_{\text{max}} \, s^n}{K_{\text{m}}^n + s^n}$$

with Hill coefficient $n = 1$, or $n > 1$ if the reaction is cooperative.

Next, let us make things more interesting by adding a degradation term $-\lambda p$.

In other words, we suppose that product is being produced, but it is also being used up or degraded, at some linear rate λp, where λ is some positive constant.

We obtain the following equation:

$$\frac{dp}{dt} = \frac{V_{\text{max}} \, s^n}{K_{\text{m}}^n + s^n} - \lambda p$$

for $p(t)$.

As far as p is concerned, this looks like an equation $\frac{dp}{dt} = \mu - \lambda p$, so as $t \to \infty$ we have that $p(t) \to \frac{\mu}{\lambda}$.

Let us take $\lambda = 1$ just to make notations easier.[34] Then the steady state obtained for p is:

$$p(\infty) \;=\; \frac{V_{max}\, s^n}{K_m^n + s^n}$$

Let us first consider the case $n = 1$.
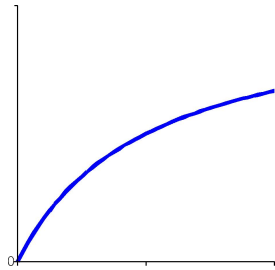
By analogy, if s would be the displacement of a slider or dial, a light-dimmer behaves in this way:

the steady-state as a function of the "input" concentration s (which we are assuming is some constant) is *graded*, in the sense that it is proportional to the parameter s (over a large range of values s; eventually, it saturates).

The case $n = 1$ gives what is called a *hyperbolic* response, in contrast to *sigmoidal* response that arises from cooperativity ($n > 1$).

As n gets larger, the plot of $\frac{V_{max}s^n}{K_m^n+s^n}$ becomes essentially a step function with a transition at $s = K_m$.

Here are plots with $V_{max} = 1$, $K_m = 0.5$, and $n = 3, 20$:
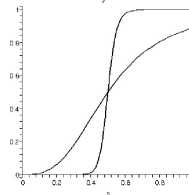
The sharp increase, and saturation, means that a value of s which is under some threshold (roughly, $s < K_m$) will not result in an appreciable result ($p \approx 0$, in steady state) while a value that is over this threshold will give an abrupt change in result ($p \approx V_{max}$, in steady state).

A "binary" response is thus produced from cooperative reactions.
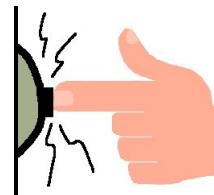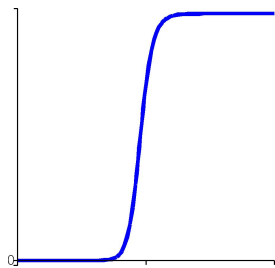
The behavior of closer to that of a doorbell: if we don't press hard enough, nothing happens; if we press with the right amount of force (or more), the bell rings.

Ultrasensitivity

Sigmoidal responses are characteristic of many signaling cascades, which display what biologists call an *ultrasensitive* response to inputs. If the purpose of a signaling pathway is to decide whether a gene

---

[34]If λ is arbitrary, just replace V_{max} by V_{max}/λ everywhere.

should be transcribed or not, depending on some external signal sensed by a cell, for instance the concentration of a ligand as compared to some default value, such a binary response is required.

Cascades of enzymatic reactions can be made to display ultrasensitive response, as long as at each step there is a Hill coefficient $n > 1$, since the derivative of a composition of functions $f_1 \circ f_2 \circ \ldots \circ f_k$ is, by the chain rule, a product of derivatives of the functions making up the composition.

Thus, the slopes get multiplied, and a steeper nonlinearity is produced. In this manner, a high effective cooperativity index may in reality represent the result of composing several reactions, perhaps taking place at a faster time scale, each of which has only a mildly nonlinear behavior.

2.7.2 Adding Positive Feedback

Next, we build up a more complicated situation by adding *feedback* to the system.

Let us suppose that the substrate concentration is not constant, but instead it depends monotonically on the product concentration.[35]

For example, the "substrate" s might represent a transcription factor which binds to DNA and instructs the production of mRNA for a protein p, and the protein p, in turn, instructs the transcription of s.

Or, possibly, $p = s$, meaning that p serves to enhance its own transcription. (autocatalytic process).

The effect of p on s may be very complex, involving several intermediaries.
However, since all we want to do here is to illustrate the main ideas, we'll simply say that $s(t) = \alpha p(t)$, for some constant α.

Therefore, the equation for p becomes now:

$$\frac{dp}{dt} = \frac{V_{\max}\,(\alpha p)^n}{K_{\mathrm{m}}^n + (\alpha p)^n} - \lambda p$$

or, if we take for simplicity[36] $\alpha = 1$ and $\lambda = 1$:

$$\frac{dp}{dt} = \frac{V_{\max}\,p^n}{K_{\mathrm{m}}^n + p^n} - p\,.$$

What are the possible steady states of this system with feedback?

Let us analyze the solutions of the differential equation, first with $n = 1$.
We plot the first term (formation rate) together with the second one (degradation):
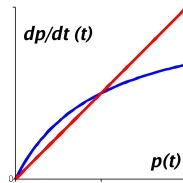
Observe that, for small p, the formation rate is larger than the degradation rate,
while, for large p, the degradation rate exceeds the formation rate.
Thus, the concentration $p(t)$ *converges to a unique intermediate value.*

---

[35]If we wanted to give a careful mathematical argument, we'd need to do a time-scale separation argument in detail. We will proceed very informally.

[36]Actually, we can always rescale p and t and rename parameters so that we have this simpler situation, anyway.

Bistability arises from sigmoidal formation rates

In the cooperative case (i.e., $n > 1$), however, the situation is far more interesting!
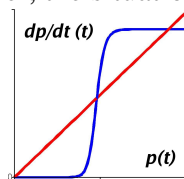
• for small p the degradation rate is larger than the formation rate, so the concentration $p(t)$ *converges to a low value*,

• but for large p the formation rate is larger than the degradation rate, and so the concentration $p(t)$ *converges to a high value* instead.

In summary, two stable states are created, one "low" and one "high", by this interaction of formation and degradation, if one of the two terms is sigmoidal.
(There is also an intermediate, unstable state.)

Instead of graphing the formation rate and degradation rate separately, one may (and we will, from now on) graph the right hand side

$$\frac{V_{\max} p^n}{K_{\mathrm{m}}^n + p^n} - p$$

as a function of p. From this, the phase line can be read-out, as done in your ODE course.

For example, here is the graph of

$$\frac{V_{\max} p^n}{K_{\mathrm{m}}^n + p^n} - p$$

with $V_{\max} = 3$, $K_{\mathrm{m}} = 1$, and $n = 2$.
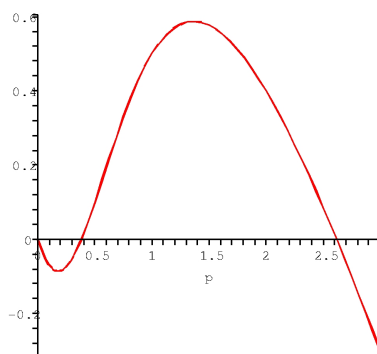
The phase line is as follows:

where $A = 0$, $B = 3/2 - 1/2 * 5^{(1/2)} \approx 0.38$, and $C = 3/2 + 1/2 * 5^{(1/2)} \approx 2.62$.

We see that A and C are stable (i.e., sinks) and the intermediate point B is a unstable (i.e., a source)

2.7.3 Cell Differentiation and Bifurcations

In unicellular organisms, cell division results in cells that are identical to each other and to the original ("mother") cell. In multicellular organisms, in contrast, cells differentiate.

Since all cells in the same organism are genetically identical, the differences among cells must result from variations of gene expression.

A central question in developmental biology is: how are these variations established and maintained?

A possible mechanism by which spatial patterns of cell differentiation could be specified during embryonic development and regeneration is based on *positional information*.[37] Cells acquire a positional value with respect to boundaries, and then use this "coordinate system" information during gene expression, to determine their fate and phenotype.
(Daughter cells inherit as "initial conditions" the gene expression pattern of the mother cells, so that a developmental history is maintained.)

In other words, the basic premise is that position in the embryo determines cell fate.
But how could this position be estimated by each individual cell?

One explanation is that there are chemicals, called *morphogens*, which are nonuniformly distributed. Typically, morphogens are RNA or proteins.
They instruct cells to express certain genes, depending on position-dependent concentrations (and slopes of concentrations, i.e. gradients).
When different cells express different genes, the cells develop into distinct parts of the organism.

An important concept is that of *polarity*: opposite ends of a whole organism or of a given tissue (or sometimes, of a single cell) are different, and this difference is due to morphogen concentration differences.

Polarity is initially determined in the embryo.
It may be established initially by the site of sperm penetration, as well as environmental factors such as gravity or pH.

The existence of morphogens and their role in development were for a long time just an elegant mathematical theory, but recent work in developmental biology has succeeded in demonstrating that embryos do in fact use morphogen gradients. This has been shown for many different species, although most of the work is done in fruit flies.[38]

Using mathematical models of morphogens and positional information, it is in principle possible to predict how mutations affect phenotype. Indeed, the equations might predict, say, that antennae in fruit flies will grow in the wrong part of the body, as a consequence of a mutation. One can then perform the actual mutation and validate the prediction by letting the mutant fly develop.

How can small differences in morphogen lead to abrupt changes in cell fate?

For simplicity, let us think of a "wormy" one-dimensional organism, but the same ideas apply to a full 3-d model.

---

[37]The idea of positional information is an old one in biology, but it was Louis Wolpert in 1971 who formalized it, see: Lewis, J., J.M. Slack, and L. Wolpert, "Thresholds in development," J. Theor. Biol. 1977, 65: 579-590.
A good, non-mathematical, review article is "One hundred years of positional information" by Louis Wolpert, appeared in *Trends in Genetics*, 1996, 12:359-64.
[38]A nice expository article (focusing on frogs) is: Jeremy Green, "Morphogen gradients, positional information, and Xenopus: Interplay of theory and experiment," *Developmental Dynamics*, 2002, 225: 392-408.

signal highest here

signal lower

cell # 1	cell # 2		cell # k		cell # N
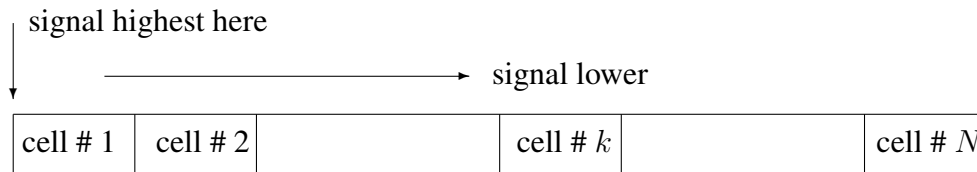
We suppose that each cell may express a protein P whose level (concentration, if you wish) "p" determines a certain phenotypical (i.e., observable) characteristic.

As a purely hypothetical and artificial example, it may be the case that P can attain two very distinct levels of expression: "very low" (or zero) or "very high," and that a cell will look like a "nose" cell if p is high, and like a "mouth" cell if p is low.[39]

Moreover, we suppose that a certain morphogen S (we use S for "signal") affects the expression mechanism for the gene for P, so that the concentration s of S in the vicinity of a particular cell influences what will happen to that particular cell.

The concentration of the signaling molecule S is supposed to be highest at the left end, and lowest at the right end, of the organism, and it varies continuously. (This may be due to the mother depositing S at one end of the egg, and S diffusing to the other end, for example.)

The main issue to understand is: since nearby cells detect only slightly different concentrations of S, how can "sudden" changes of level of P occur?

$s = 1$	$s = 0.9$	$s = 0.8$	$s = 0.7$	$s = 0.6$	$s = 0.5$	$s = 0.4$	$s = 0.3$	$s = 0.2$
$p \approx 1$	$p \approx 1$	$p \approx 1$	$p \approx 1$	$p \approx 1$	$p \approx 0$	$p \approx 0$	$p \approx 0$	$p \approx 0$
nose cell	nose cell	nose cell	nose cell	nose cell	mouth cell	mouth cell	mouth cell	mouth cell

In other words, why don't we find, in-between cells that are part of the "nose" (high p) and cells that are part of the "mouth" (low p), cells that are, say, "3/4 nose, 1/4 mouth"?

We want to understand how this "thresholding effect" could arise.

The fact that the DNA in all cells of an organism is, in principle, identical, is translated mathematically into the statement that all cells are described by the same system of equations, but we include an input parameter in these equations to represent the concentration s of the morphogen near any given cell.[40]

In other words, we'll think of the evolution on time of chemicals (such as the concentration of the protein P) as given by a differential equation:

$$\frac{dp}{dt} = f(p, s)$$

(of course, realistic models contain many proteins or other substances, interacting with each other through mechanisms such as control of gene expression and signaling; we use an unrealistic single equation just to illustrate the basic principle).

We assume that from each given initial condition $p(0)$, the solution $p(t)$ will settle to some steady state $p(\infty)$; the value $p(\infty)$ describes what the level of P will be after a transient period. We think of $p(\infty)$ as determining whether we have a "nose-cell" or a "mouth-cell."

---

[39]Of course, a real nose has different types of cells in it, but for this silly example, we'll just suppose that they all look the same, but they look very different from mouth-like cells, which we also assume all look the same.

[40]We assume, for simplicity, that s constant for each cell, or maybe the cell samples the average value of s around the cell.

Of course, $p(\infty)$ depends on the initial state $p(0)$ as well as on the value of the parameter s that the particular cell measures.

We will assume that, at the start of the process, all cells are in the same initial state $p(0)$.

So, we need that $p(\infty)$ be drastically different only due to a change in the parameter s.[41]

To design a realistic "f," we start with the positive feedback system that we had earlier used to illustrate bi-stability, and we add a term "$+ks$" as the simplest possible mechanism by which the concentration of signaling molecule may influence the system.[42]:

$$\frac{dp}{dt} \;=\; f(p,s) \;=\; \frac{V_{\max}\, p^n}{K_m^n + p^n} \;-\; \lambda p \;+\; ks\,.$$

Let us take, to be concrete, $k=5$, $V_{\max}=15$, $\lambda=7$, $K_m=1$, Hill coefficient $n=2$, and $\alpha=1$.

There follow the plots of $f(p,s)$ versus p, for three values of s:

$$s < s^*, \quad s = s^*, \quad s > s^*, \quad \text{where } \; s^* \approx .268\,.$$
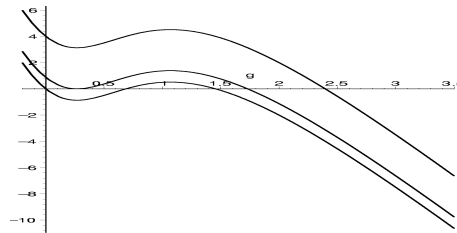
The respective phase lines are now shown below the graphs:
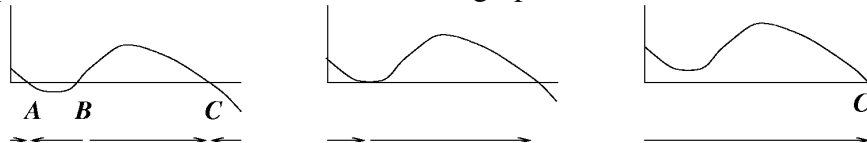
We see that for $s < s^*$, there are two sinks (stable steady states), marked A and C respectively, as well as a source (unstable steady state), marked B.

We think of A as the steady state protein concentration $p(\infty)$ representing mouth-like cells, and C as that for nose-like cells.

Of course, the exact position of A depends on the precise value of s. Increasing s by a small amount means that the plot moves up a little, which means that A moves slightly to the right. Similarly, B moves to the left and C to the right.

However, we may still think of a "low" and a "high" stable steady state (and an "intermediate" unstable state) in a qualitative sense.

Note that B, being an unstable state, will never be found in practice: the smallest perturbation makes the solution flow away from it.

For $s > s^*$, there is only one steady state, which is stable. We denote this state as C, because it corresponds to a high concentration level of P.

---

[41] This is the phenomenon of "bifurcations," which you should have encountered in the previous differential equations course.

[42] This term could represent the role of s as a transcription factor for p. The model that we are considering is the one proposed in the original paper by Lewis et al.

Once again, the precise value of C depends on the precise value of s, but it is still true that C represents a "high" concentration.

Incidentally, a value of s *exactly* equal to s^* will never be sensed by a cell: there is zero probability to have this precise value.

Now, assume that all cells in the organism start with no protein, that is, $p(0) = 0$.

The left-most cells, having $s > s^*$, will settle into the "high state" C, i.e., they will become nose-like.

The right-most cells, having $s < s^*$, will settle into the "low state" A, i.e., they will become mouth-like.

So we see how a sharp transition between cell types is achieved, merely due to a change from $s > s^*$ to $s < s^*$ as we consider cells from the left to the right end of the organism.

$s > s^*$	$s > s^*$	$s > s^*$	$s > s^*$	$s > s^*$	$s < s^*$	$s < s^*$	$s < s^*$	$s < s^*$
$p \approx C$ nose cell	$p \approx C$ nose cell	$p \approx C$ nose cell	$p \approx C$ nose cell	$p \approx C$ nose cell	$p \approx A$ mouth cell	$p \approx A$ mouth cell	$p \approx A$ mouth cell	$p \approx A$ mouth cell

Moreover, this model has a most amazing feature, which corresponds to the fact that, once a cell's fate is determined, it will not revert[43] to the original state.

Indeed, suppose that, after a cell has settled to its steady state (high or low), we now suddenly "wash-out" the morphogen, i.e., we set s to a very low value.

The behavior of *every* cell will now be determined by the phase line for low s:
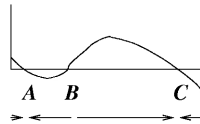
This means that any cell starting with "low" protein P will stay low, and any cell starting with "high" protein P will stay high.

A permanent memory of the morphogen effect is thus imprinted in the system, even after the signal is "turned-off"!

A little exercise to test understanding of these ideas.

A multicellular 1-d organism as before is considered. Each cell expresses a certain gene X according to the same differential equation

$$\frac{dx}{dt} = f(x) + a.$$

The cells at the left end receive a low signal a, while those at the right end see a high signal a (and the signal changes continuously in between).
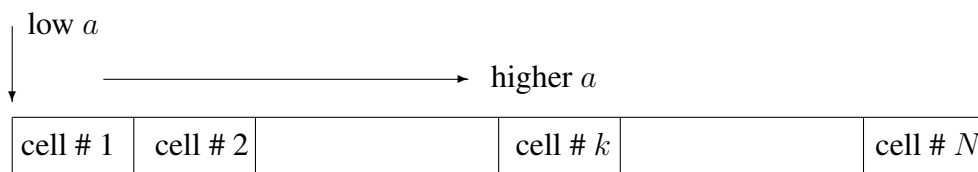
---
[43] As with stem cells differentiating into different tissue types.

The following plots show the graph of $f(x) + a$, for small, intermediate, and large a respectively.
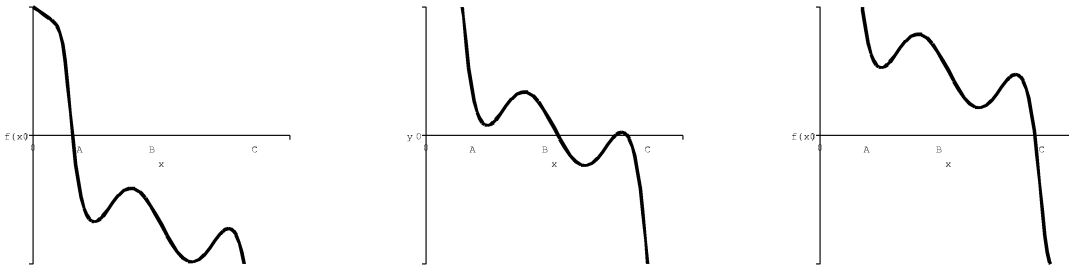
We indicate a roughly "low" level of x by the letter "A," an "intermediate" level by "B," and a "high" level by "C."

Question: Suppose that the level of expression starts at $x(0) = 0$ for every cell.

(1) What pattern do we see after things settle to steady state?

(2) Next suppose that, *after* the system has so settled, we now suddenly change the level of the signal a so that now *every* cell sees the *same* value of a. This value of a that every cell is exposed to, corresponds to this plot of $f(x) + a$:
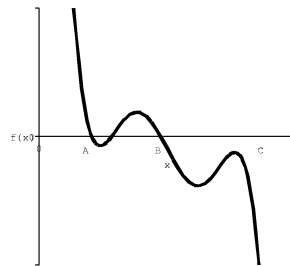
What pattern will the organism settle upon?

Answer:

Let us use this picture:

left cell	left cell	left cell	center cell	center cell	center cell	right cell	right cell	right cell

- Those cells located toward the left will see these "instructions of what speed to move at:"
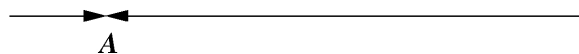
Therefore, starting from $x = 0$, they settle at a "low" gene expression level, roughly indicated by A.

- Cells around the center will see these "instructions:"

(There is an un-labeled unstable state in-between B and C.) Thus, starting from $x = 0$, they settle at an "intermediate" level B.

- Finally, those cells toward the left will see these "instructions:"
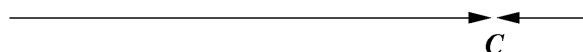
Therefore, starting from $x = 0$, they will settle at a "high" level C.

In summary, the pattern that we observe is:

$$AAABBBCCC.$$

(There may be many A's, etc., depending on how many cells there are, and what exactly is the graph of f. We displayed 3 of each just to give the general idea!)

Next, we suddenly "change the rules of the game" and ask them *all* to follow these instructions:
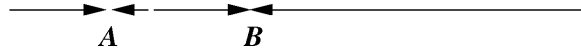
(There is an un-labeled unstable state in-between A and B.) Now, cells that started (from the previous stage of our experiment) near enough A will approach A, cells that were near B approach B, and cells that were near C have "their floor removed from under them" so to speak, and they are being now told to move left, i.e. all the way down to B.

In summary, we have that starting at $x = 0$ at time zero, the pattern observed after the first part of the experiment is:

$$AAABBBCCC,$$

and after the second part of the experiment we obtain this final configuration:

$$AAABBBBBB.$$

(Exactly how many A's and B's depends on the precise form of the function f, etc. We are just representing the general pattern.)

2.7.4 Sigmoidal responses without cooperativity: Goldbeter-Koshland

Highly sigmoidal responses require a large Hill coefficient n_H. In 1981, Goldbeter and Koshland made a simple but strikingly interesting and influential observation: one can obtain such responses even without cooperativity. The starting point is a reaction such as the "futile cycle" (2.10):

$$E + P \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} C \xrightarrow{k_2} E + Q, \qquad F + Q \underset{k_{-3}}{\overset{k_3}{\rightleftarrows}} D \xrightarrow{k_4} F + P.$$
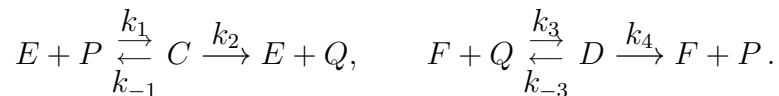
To simplify, we take a quasi-steady state approximation, so that (using lower case letters for concentrations), $dq/dt = \frac{V_{\max} ep}{K + p} - \frac{\widetilde{V_{\max}} fq}{L + q}$ and $dp/dt = -dq/dt$. Thus $p + q$ is constant, and by picking appropriate units we let $q = 1 - p$ (so that p, q are now the fractions of unmodified substrate, respectively). Writing "x" instead of "p", we have that steady states should be solutions of

$$r \frac{x}{K + x} = \frac{1 - x}{L + 1 - x}, \tag{2.18}$$

where $r := (V_{\max} / \widetilde{V_{\max}})(e/f)$ is proportional to the ratio of the concentrations of the two enzymes. Sketching the two curves to find the intersection points, we see that for small K, L ("zero order" regime, in the sense that the production rate is approximately constant, except for p, q very small), there is a very sharp dependence of the steady state x on the value of r, changing from $x \approx 1$ to $x \approx 0$

at $r = 1$. In contrast, for K and L large ("first order" or almost linear, regime) the dependence is far smoother.

Left:

dotted lines: graphs of $\frac{1-x}{L+1-x}$

solid lines: graphs of $r\frac{x}{K+x}$

for $r < 1$ (left) and $r > 1$ (right)

top sketches: $K, L \ll 1$, bottom: large K, L

Right: dependence $x = G(r)$
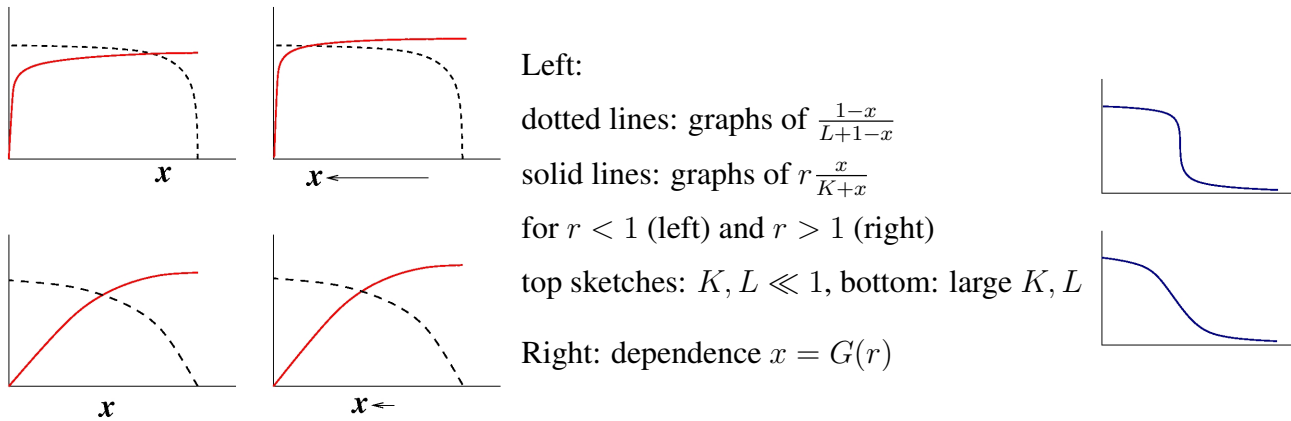
In summary, for small K, L, we have a very sigmoidal response, with no need for cooperativity.

Solving $x = G(r)$, G is the "GoldbeterKoshland function".

2.7.5 Bistability with two species

There is a molecular-level analog of the "species competition" models in ecology, as follows.

Suppose that we have two genes that code for the proteins X and Y. In order to get terms that represent an analog of "logistic growth," we need to have expressions of the form "$x(A - x)$" in the growth rates. Let us assume that there is positive feedback of each of X and Y on themselves (self-induction), but that the dimers XX, XY, YY act as repressors. Using lower case letters for concentrations, we are led to equations as follows:

$$\frac{dx}{dt} = \mu_1 x - \alpha_1 x^2 - \gamma_{12} xy$$
$$\frac{dy}{dt} = \mu_2 y - \alpha_2 y^2 - \gamma_{21} xy \,.$$

Several different behaviors result, depending on the values of the parameters, just as with population models. Let us take the case: $\alpha_2/\gamma_{12} < \mu_1/\mu_2 < \gamma_{21}/\alpha_1$. There are four steady states, but the behavior of the trajectories depends strongly on the initial concentrations: we always converge to either $x(\infty) = 0$ or $y(\infty) = 0$, *except* when we start exactly on the stable manifold of the saddle point ("principle of competitive exclusion").
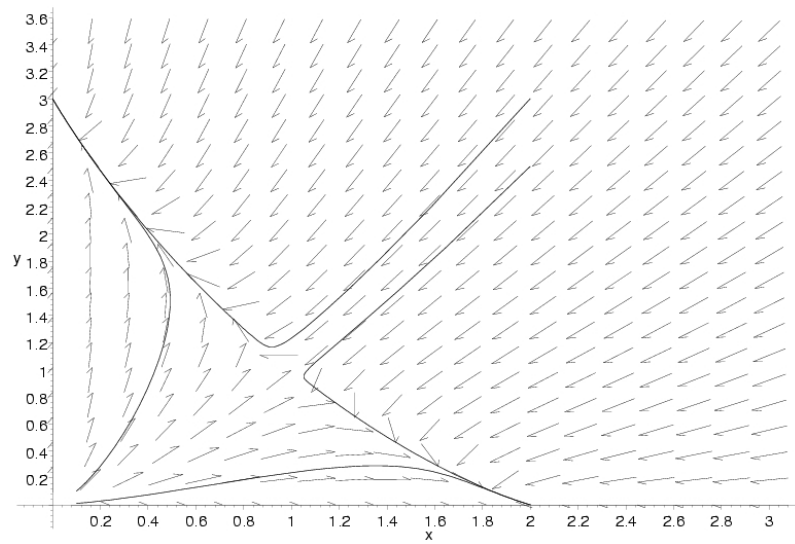
2.8   Periodic Behavior

Periodic behaviors (i.e, oscillations) are very important in biology, appearing in diverse areas such as neural signaling, circadian rythms, and heart beats.

You have seen examples of periodic behavior in the differential equations course, most probably the harmonic oscillator (mass spring system with no damping)

$$\frac{dx}{dt} = y$$
$$\frac{dy}{dt} = -x$$

whose trajectories are circles, or, more generally, linear systems with eigenvalues that are purely imaginary, leading to ellipsoidal trajectories:
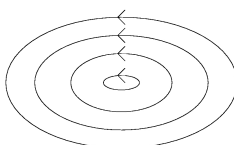
A serious limitation of such linear oscillators is that they are not *robust*:

Suppose that there is a small perturbation in the equations:

$$\frac{dx}{dt} = y$$
$$\frac{dy}{dt} = -x + \varepsilon y$$

where $\varepsilon \neq 0$ is small. The trajectories are not periodic anymore!

Now dy/dt doesn't balance dx/dt just right, so the trajectory doesn't "close" on itself:
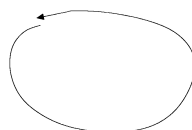
Depending on the sign of ε, we get a stable or an unstable spiral.

When dealing with electrical or mechanical systems, it is often possible to construct things with precise components and low error tolerance. In biology, in contrast, things are too "messy" and oscillators, if they are to be reliable, must be more "robust" than simple harmonic oscillators.

Another disadvantage of simple linear oscillations is that if, for some reason, the state "jumps" to another position[44] then the system will simply start oscillating along a different orbit and never come back to the original trajectory:
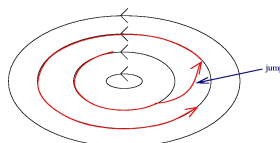
To put it in different terms, the particular oscillation depends on the initial conditions. Biological objects, in contrast, tend to reset themselves (e.g., your internal clock adjusting after jetlag).

---

[44] the "jump" is not described by the differential equation; think of the effect of some external disturbance that gives a "kick" to the system

2.8.1 Periodic Orbits and Limit Cycles

A *(stable) limit cycle* is a periodic trajectory which attracts other solutions (at least those starting nearby) to it.[45]
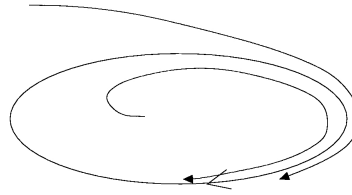
Thus, a member of a family of "parallel" periodic solutions (as for linear centers) is *not* called a limit cycle, because other close-by trajectories remain at a fixed distance away, and do not converge to it.

Limit cyles are "robust" in ways that linear periodic solutions are not:

• If a (small) perturbation moves the state to a different initial state away from the cycle, the system will return to the cycle by itself.

• If the dynamics changes a little, a limit cycle will still exist, close to the original one.

The first property is obvious from the definition of limit cycle. The second property is not very difficult to prove either, using a "Lyapunov function" argument. (Idea sketched in class.)

2.8.2 An Example of Limit Cycle

In order to understand the definition, and to have an example that we can use for various purposes later, we will consider the following system[46]:

$$\begin{aligned} dx_1/dt &= \mu x_1 - \omega x_2 + \theta x_1(x_1^2 + x_2^2) \\ dx_2/dt &= \omega x_1 + \mu x_2 + \theta x_2(x_1^2 + x_2^2). \end{aligned}$$

where we pick $\theta = -1$ for definiteness, so that the system is:

$$\begin{aligned} dx_1/dt &= \mu x_1 - \omega x_2 - x_1(x_1^2 + x_2^2) \\ dx_2/dt &= \omega x_1 + \mu x_2 - x_2(x_1^2 + x_2^2). \end{aligned}$$

(Note that if picked $\theta = 0$, we would have a linear harmonic oscillator, which has no limit cycles.)

There are two other ways to write this system which help us understand it better.

The first is to use polar coordinates.

We let $x_1 = \rho \cos \varphi$ and $x_2 = \rho \sin \varphi$, and differentiate with respect to time. Equating terms, we obtain separate equations for the magnitude ρ and the argument φ, as follows:

$$\begin{aligned} d\rho/dt &= \rho(\mu - \rho^2) \\ d\varphi/dt &= \omega. \end{aligned}$$

(The transformation into polar coordinates is only valid for $x \neq 0$, that is, if $\rho > 0$, but the transformed equation is formally valid for all ρ, φ.)

---

[45]Stable limit cycles are to all periodic trajectories as stable steady states are to all steady states.
[46]of course, this is a purely mathematical example

Another useful way to rewrite the system is in terms of complex numbers; a problem asks for that.

We now analyze the system using polar coordinates.

Since the differential equations for ρ and φ are decoupled, we may analyze each of them separately.

The φ-equation $d\varphi/dt = \omega$ tells us that the solutions must be rotating at speed ω (counter-clockwise, if $\omega > 0$).

Let us look next at the scalar differential equation $d\rho/dt = \rho(\mu - \rho^2)$ for the magnitude r.

When $\mu \leq 0$, the origin is the only steady state, and every solution converges to zero. *This means that the full planar system is so that all trajectories spiral into the origin.*

When $\mu > 0$, the origin of the scalar differential equation $d\rho/dt = \rho(\mu - \rho^2)$ becomes unstable[47], as we can see from the phase line. In fact, the velocity is negative for $\rho > \sqrt{\mu}$ and positive for $\rho < \sqrt{\mu}$, so that there is a sink at $\rho = \sqrt{\mu}$. *This means that the full planar system is so that all trajectories spiral into the circle of radius $\sqrt{\mu}$, which is, therefore, a limit cycle.*
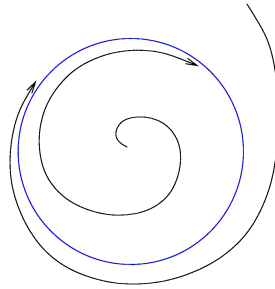
(Expressed in terms of complex-numbers, $z(t) = \sqrt{\mu}e^{i\omega t}$ is the limit cycle.)

Note that the oscillation has magnitude $\sqrt{\mu}$ and frequency ω.

Unfortunately, it is quite difficult to actually prove that a limit cycle exists, for more general systems.

But for systems of **two** equations, there is a very powerful criterion.

2.8.3 Poincaré-Bendixson Theorem

Suppose *a bounded region D in the plane is so that* **no trajectories can exit** D,

(in other words, we have a "forward-invariant" or "trapping" region)

and either that there are **no steady states** *inside, or there is a single steady state that is repelling.*[48]

Then, *there is a periodic orbit inside D.*

---

[47]the passage from $\mu < 0$ to $\mu > 0$ is a typical example of what is called a "supercritical Hopf bifurcation"

[48]Looking at the trace/determinant plane, we see that repelling points are those for which both the determinant and the trace of the Jacobian are positive, since the other quadrants represent either stable points or saddles.
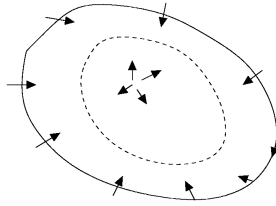
This theorem is proved in advanced differential equations books; the basic idea is easy to understand: if we start near the boundary, we must go towards the inside, and cannot cross back (because trajectories cannot cross). Since it cannot approach a source, the trajectory must approach a periodic orbit. (Idea sketched in class.)

We gave a simple version, sufficient for our purposes; one can state the theorem a little more generally, saying that all trajectories will converge to either steady states, limit cycles, or "connections" among steady states. One such version is as follows: if the omega-limit set $\omega(x)$[49] of a trajectory is compact, connected, and contains only finitely many equilibria, then these are the only possibilities for $\omega(x)$:

- $\omega(x)$ is a steady state, or

- $\omega(x)$ is a periodic orbit, or

- $\omega(x)$ is a homoclinic or heteroclinic connection.

It is also possible to prove that if there is a unique periodic orbit, then it must be a limit cycle.

In general, finding an appropriate region D is usually quite hard; often one uses plots of solutions and/or nullclines in order to guess a region.

Invariance of a region D can be checked by using the following test: the outward-pointing normal vectors, at any point of the boundary of D, must make an angle of at least 90 degrees with the vector field at that point. Algebraically, this means that the dot product must be ≤ 0 between a normal $\vec{n}$ and the vector field:

$$\left(\frac{dx}{dt}, \frac{dy}{dt}\right) \cdot \vec{n} \leq 0$$

at any boundary point.[50]
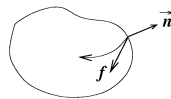
Let us work out the example:

$$dx/dt_1 = \mu x_1 - \omega x_2 - x_1(x_1^2 + x_2^2)$$
$$dx/dt_2 = \omega x_1 + \mu x_2 - x_2(x_1^2 + x_2^2)$$

with $\mu > 0$, using P-B. (Of course, we already know that the circle with radius $\sqrt{\mu}$ is a limit cycle, since we showed this by using polar coordinates.)

---

[49]This is the set of limit points of the solution starting from an initial condition x

[50]If the dot product is strictly negative, this is fairly obvious, since the vector field must then "point to the inside" of D. When the vectors are exactly perpendicular, the situation is a little more subtle, especially if there are corners in the boundary of D (what is a "normal" at a corner?), but the equivalence is still true. The mathematical field of "nonsmooth analysis" studies such problems of invariance, especially for regions with possible corners.

We must find a suitable invariant region, one that contains the periodic orbit that we want to show exists. Cheating (because if we already know it is there, we don't need to find it!), we take as our region D the disk with radius $\sqrt{2\mu}$. (Any large enough disk would have done the trick.)

To show that D is a trapping region, we must look at its boundary, which is the circle of radius $\sqrt{2\mu}$, and show that the normal vectors, at any point of the boundary, form an angle of at least 90 degrees with the vector field at that point. This is exactly the same as showing that the dot product between the normal and the vector field is negative (or zero, if tangent).

At any point on the circle $x_1^2 + x_2^2 = 2\mu$, a normal vector is (x_1, x_2) (since the arrow from the origin to the point is perpendicular to circle), and the dot product is:

$$[\mu x_1 - \omega x_2 - x_1(x_1^2 + x_2^2)]\, x_1 + [\omega x_1 + \mu x_2 - x_2(x_1^2 + x_2^2)]\, x_2 \;=\; (\mu - (x_1^2 + x_2^2))(x_1^2 + x_2^2) \;=\; -2\mu^2 \;<\; 0\,.$$

Thus, the vector field points inside and the disk of radius $2\sqrt{\mu}$ is a trapping region.
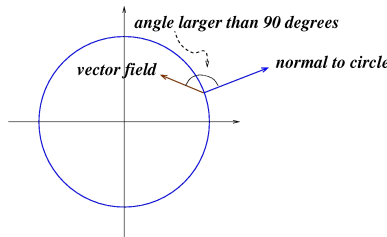
The only steady state is $(0,0)$, which we can see by noticing that if $\mu x_1 - \omega x_2 - x_1(x_1^2 + x_2^2) = 0$ and $\omega x_1 + \mu x_2 - x_2(x_1^2 + x_2^2) = 0$ then multiplying by x_1 the first equation, and the second by x_2, we obtain that $(\mu + x_1^2 + x_2^2)(x_1^2 + x_2^2) = 0$, so $x_1 = x_2 = 0$.

Linearizing at the origin, we have an unstable spiral. (Homework: check!) Thus, the only steady state is repelling, which is the other property that we needed. So, we can apply the P-B Theorem.

We conclude that there is a periodic orbit inside this disk.[51]

2.8.4   The Van der Pol Oscillator

A typical way in which periodic orbits arise in models in biology and many other fields can be illustrated with the well-known *Van der Pol oscillator*.[52] After some changes of variables, which we do not discuss here, the van der Pol oscillator becomes this system:

$$\frac{dx}{dt} = y + x - \frac{x^3}{3}$$
$$\frac{dy}{dt} = -x$$

The only steady state is at $(0,0)$, which repels, since the Jacobian has positive determinant and trace:

$$\left. \begin{pmatrix} 1 - x^2 & 1 \\ -1 & 0 \end{pmatrix} \right|_{(0,0)} = \begin{pmatrix} 1 & 1 \\ -1 & 0 \end{pmatrix}\,.$$

---

[51]In fact, using annular regions $\sqrt{\mu} - \varepsilon < x_1^2 + x_2^2 < \sqrt{\mu} + \varepsilon$, one can prove by a similar argument that the periodic orbit is unique, and, therefore, is a limit cycle.

[52]Balthazar van der Pol was a Dutch electrical engineer, whose oscillator models of vacuum tubes are a routine example in the theory of limit cycles; his work was motivated in part by models of the human heart and an interest in arrhythmias. The original paper was: B. van der Pol and J. van der Mark, *The heartbeat considered as a relaxation oscillation, and an electrical model of the heart*, Phil. Mag. Suppl. #6 (1928), pp. 763–775.

We will show that there are periodic orbits (one can also show there is a limit cycle, but we will not do so), by applying Poincaré-Bendixson.

To apply P-B, we consider the following special region:
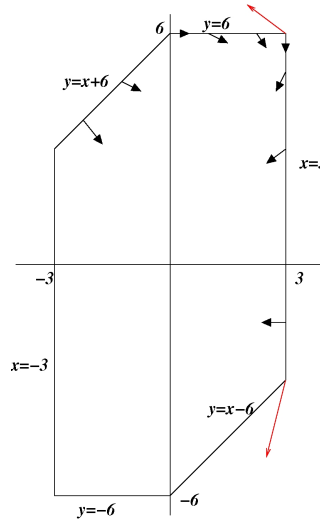
We will prove that, on the boundary, the vector field point inside, as shown by the arrows.
The boundary is made up of 6 segments, but, by symmetry,
(since the region is symmetric and the equation is odd), it is enough to consider 3 segments:

$$x = 3,\ -3 \le y \le 6 \qquad\qquad y = 6,\ 0 \le x \le 3 \qquad\qquad y = x + 6,\ -3 \le x \le 0\,.$$

$\underline{x = 3,\ -3 \le y \le 6}$:
we may pick $\vec{\nu} = (1, 0)$, so $\left(\frac{dx}{dt}, \frac{dy}{dt}\right) \cdot \vec{n} = \frac{dx}{dt}$ and, substituting $x = 3$ into $y + x - \frac{x^3}{3}$, we obtain:

$$\frac{dx}{dt} = y - 6 \le 0\,.$$

Therefore, we know that the vector field points to the left, on this boundary segment.

We still need to make sure that things do not "escape" through a corner, though. In other words, we need to check that, on the corners, there cannot be any arrows as the red ones.

Actually, one can prove that, for convex regions built up piecewise as in all our examples, it is enough to verify the normal vector condition at non-corner points. (So, you can skip verifications of corners in all homework problems.) However, for expository reasons, we show now how to verify the conditions at corners without appealing to that (unproved here) fact.

At the top corner, $x = 3, y = 6$, we have $dy/dt = -3 < 0$, so that the corner arrow must point down, and hence "SW", so we are OK. At the bottom corner, also $dy/dt = -3 < 0$, and $dx/dt = -9$, so the vector field at that point also points inside.

$\underline{y = 6,\ 0 \le x \le 3}$:
we may pick $\vec{\nu} = (0, 1)$, so

$$\left(\frac{dx}{dt}, \frac{dy}{dt}\right) \cdot \vec{n} = \frac{dy}{dt} = -x \le 0\,,$$

and corners are also OK (for example, at $(0, 6)$: $dx/dt = 6 > 0$).

$\underline{y = x + 6,\ -3 \le x \le 0}$:

We pick the outward normal $\vec{\nu} = (-1, 1)$ and take dot product:

$$\begin{pmatrix} y + x - x^3/3 \\ -x \end{pmatrix} \cdot \begin{pmatrix} -1 \\ 1 \end{pmatrix} = -2x - y + x^3/3 \,.$$

Evaluated at $y = x + 6$, this is:

$$\frac{x^3}{3} - 3x - 6, \ -3 \le x \le 0 \,.$$

The function $f(x) = \frac{x^3}{3} - 3x - 6$ has value -6 at both endpoints $x = -3$ and $x = 0$ of the interval $-3 \le x \le 0$, and has a local maximum at $x = -\sqrt{3}$. At this local maximum, it has the value $-\sqrt{3} + 3\sqrt{3} - 6 = 2\sqrt{3} - 6 < 0$. Thus, the dot product is negative. (One can also check corners.)

2.8.5 Bendixson's Criterion

There is a useful criterion to help conclude *there cannot be any* periodic orbit in a given a simply-connected (no holes) region D:

If *the divergence of the vector field is everywhere positive*[53] *or is everywhere negative inside D,* **then** *there cannot be a periodic orbit inside D.*

Sketch of proof (by contradiction):

Suppose that there is some such periodic orbit, which describes a simple closed curve C. We replace the original D by the inside component of C; this set D is again simply-connected.

Recall that the divergence of $F(x, y) = \begin{pmatrix} f(x, y) \\ g(x, y) \end{pmatrix}$ is defined as:

$$\frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} \,.$$

The Gauss Divergence Theorem (or "Green's Theorem") says that:

$$\int\int_D \operatorname{div} F(x, y)\, dxdy \ = \ \int_C \vec{n} \cdot F$$

(the right-hand expression is the line integral of the dot product of a unit outward normal with F).[54]

Now, saying that C is an orbit means that F is tangent to C, so the dot product is zero, and therefore

$$\int\int_D \operatorname{div} F(x, y)\, dxdy \ = \ 0 \,.$$

But, if $\operatorname{div} F(x, y)$ is everywhere positive, then the integral is positive, and we get a contradiction. Similarly if it is everywhere negative.

Example: $dx/dt = x$, $dy/dt = y$. Here the divergence is $= 2$ everywhere, so there cannot exist any periodic orbits (inside any region).

---

[53]To be precise, everywhere nonnegative but not everywhere zero

[54]The one-dimensional analog of this is the Fundamental Theorem of Calculus: the integral of F' (which is the divergence, when there is only one variable) over an interval $[a, b]$ is equal to the integral over the boundary $\{a, b\}$ of $[a, b]$, that is, $F(b) - F(a)$.

It is very important to realize what the theorem *does not* say:

Suppose that we take the example $dx/dt = x$, $dy/dt = -y$. Since the divergence is identically zero, the Bendixson criterion tells us nothing. In fact, this is a linear saddle, so we know (for other reasons) that there are no periodic obits.

On the other hand, for the example $dx/dt = y$, $dy/dt = -x$, which also has divergence identically zero, periodic orbits exist!

2.9   Bifurcations

Let us now look a bit more closely at the general idea of bifurcations.[55]

2.9.1   How can stability be lost?

The only way in which a change of behavior can occur is if the *Jacobian is degenerate* at the given steady state. Indeed, consider a system with parameter μ:

$$dx/dt = f(x, \mu).$$

If $f_x(x^*, \mu^*)$ has no eigenvalues on the imaginary axis, then it is, in particular, nonsingular.

In that case, by the Implicit Function Theorem[56], there will be a unique steady state $x = x(\mu)$ near x^*. Moreover, since the eigenvalues depend continuously on the parameters it follows that the local behavior at such $x = x(\mu)$ is the same as that at x^*, if μ is close to μ^*.

Thus, asking that the Jacobian $f_x(x^*, \mu^*)$ be degenerate is a necessary condition that one should check when looking for bifurcation points.

At points with degenerate Jacobian, the Hessian (matrix of second derivatives) is generically nonsingular (in the sense that more constraints on parameters defining the system, or on the form of f itself, are needed in order to have a singular Hessian). Thus one talks of "generic" bifurcations.

These are the generic "codimension 1" (i.e., those obtained by varying one parameter) bifurcations for equilibria:

- *one real eigenvalue crosses at zero (saddle-node, turning point, or fold)*

  two equilibria formed (or dissappear), saddle and node

- *pair of complex eigenvalues crosses imaginary axis:*

  periodic orbits arise from Poincaré-Andronov-Hopf bifurcations

---
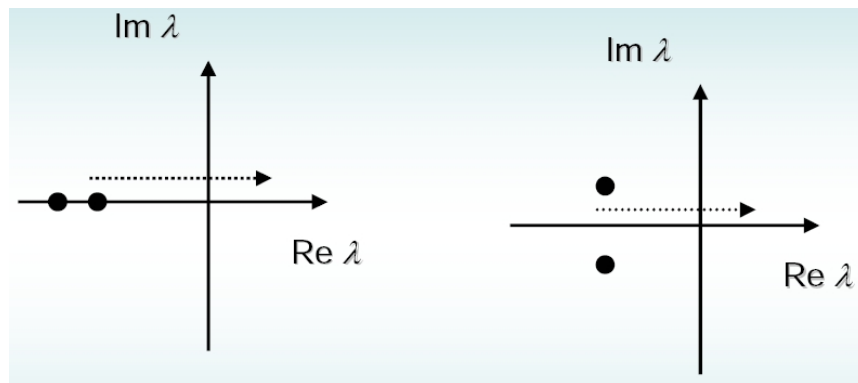
[55] Suggested references: Steven Strogatz, Nonlinear Dynamics and Chaos. Perseus Publishing, 2000. ISBN 0-7382-0453-6 and the excellent article: John D. Crawford, Introduction to bifurcation theory, Rev. Mod. Phys. 63, pp. 991-1037, 1991

[56] The IFT can be proved as follows. We need to find a function $\varphi(\mu)$ such that $f(\varphi(\mu), \mu) = 0$ for all μ in a neighborhood of μ^*. Taking total derivatives this says that $d\varphi(\mu)/d\mu = f_x(\varphi(\mu), \mu)^{-1} f_\mu(\varphi(\mu), \mu)$. As $f_x(x^*, \mu^*)$ is nonsingular, the right-hand side is well-defined on a neighborhood of (x^*, μ^*), and the existence throrem for ODE's provides a (unique) φ. For a reference to continuous dependence of eigenvalues on parameters see for example E. Sontag, Mathematical Control Theory, Springer-Verlag, 1998, Appendix A.

2.9.2   One real eigenvalue moves

Saddle-node bifurcation

For simplicity, let's assume that we have a one-dimensional system $dx/dt = f(x, \mu)$.[57] After a coordinate translation if necessary, we assume that the point of interest is $\mu^* = 0$, $x^* = 0$. We now perform a Taylor expansion, using that, at a steady state, we have $f(0, 0) = 0$, and that a bifurcation can only happen if $f_x(0, 0) = 0$:

$$dx/dt \;=\; f(x, \mu) \;=\; \mu f_\mu(0,0) \;+\; \frac{1}{2} f_{xx}(0,0) x^2 \;+\; \frac{1}{2} f_{\mu\mu}(0,0)\mu^2 \;+\; f_{x\mu}(0,0)x\mu \;+\; o(x, \mu)\,.$$

The terms that contain at least one power of μ can be collected:

$$
\begin{aligned}
f_\mu(0,0)\mu \;+\; \frac{1}{2} f_{x\mu}(0,0)x\mu \;+\; f_{\mu\mu}(0,0)\mu^2 + \ldots \;&=\; [f_\mu(0,0) + \frac{1}{2} f_{x\mu}(0,0)x + f_{\mu\mu}(0,0)\mu \ldots]\mu \\
&\approx\; f_\mu(0,0)\mu
\end{aligned}
$$

where the approximation makes sense provided that $f_\mu(0,0) \neq 0$ (since $(1/2)f_{x\mu}(0,0)x + f_{\mu\mu}(0,0)\mu \ll f_\mu(0,0)$ for small x, and similarly for higher-order terms). In the same way, we may collect all terms with at least a factor x^2, provided that $f_{xx}(0,0) \neq 0$. The conditions "$f_{xx}(0,0) \neq 0$ and $f_\mu(0,0) \neq 0$" are "generic" in the sense that, in the absence of more information (beyond the requirement that we have an equilibrium at which a bifurcation happens), they are reasonable for "random" choices of f. There results an approximation of the form $dx/dt \approx a\mu + b\xi^2$. Rescaling μ (multiplying it by a positive or negative number), we may assume that $a = 1$. Moreover, rescaling time, we may also assume that $b = \pm 1$, leading thus to the normal form:

$$dx/dt \;=\; \mu \pm x^2$$

We argued that this equation is approximately valid, under genericity conditions, but in fact it is possible to obtain it exactly, under appropriate changes of variables (Poincaré-Birkhoff theory of normal forms).

In phase-space, there will be a transition from no steady states to two steady states, as μ increases or (depending on the sign of b) decreases. Hence the alternative name "blue sky bifurcation".
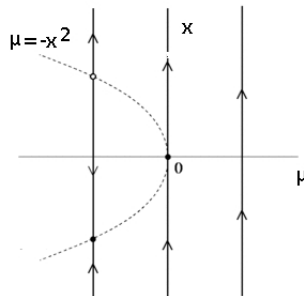
To understand the analog of this in more dimensions, suppose that we add a second equation $dy/dt = \pm y$. The pictures as we move through a bifurcation point are now as follows, assuming the above normal form:

---

[57]The theory of *Center manifolds* allows one to always reduce to this case.
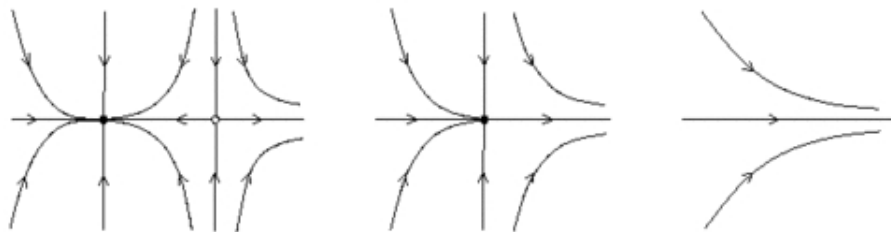
The name "saddle-node" is clear from this picture.

Transcritical bifurcation

The saddle node bifurcation is generic, like we said, assuming no requirements beyond having an equilibrium at which a bifurcation happens (and the eigenvalue on the imaginary axis being real), Often, one has additional information. For example, in a one-dimensional population model $dx/dt = k(B - x)x$ with known carrying capacity of the environment B but unknown reproduction constant k, we know that $x = B$ is an equilibrium, no matter what the value of k is. In general, if we impose the requirement that $x^* = 0$ is an equilibrium for every value of μ, $f(0, \mu) = 0$ for all μ, this implies that $\frac{\partial^k f}{\partial \mu^k}(0,0) = 0$ for all k, and thus the linear term in μ no longer dominates in the above Taylor expansion. Now we need to use the mixed quadratic term to collect all higher-order monomials. and the normal form is

$$dx/dt \;=\; \mu x - x^2$$

(precisely as with the logistic equation).

transcritical bifurcation
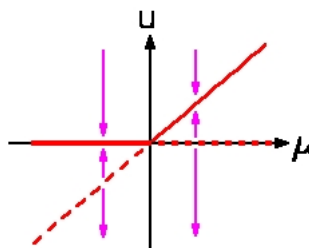
Pitchfork bifurcations

Another type of information usually available is given by symmetries imposed by physical consider-ations, For example, suppose that we know that $f(-x, \mu) = -f(x, \mu)$ (Z_2 symmetry) for all μ and x. In this case, the quadratic term vanishes, and one is led to the normal form $dx/dt = \mu x \pm x^3$ (super- and sub-critical cases):

$$dx/dt = (\mu - x^2)x \qquad\qquad dx/dt = (\mu + x^2)x$$

pitchfork bifurcation
supercritical   subcritical
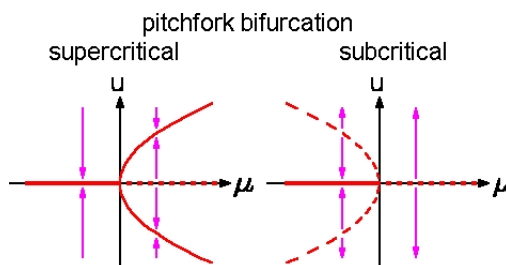
(The Hopf bifurcation, to be studied next, is closely related, though $x < 0$ does not play a rule in that case, since "x" will correspond to the norm of a point in the plane.)

2.9.3   Hopf Bifurcations

Mathematically, periodic orbits often arise from the *Hopf Bifurcation* phenomenon.

The Hopf (or "Poincaré-Andronov-Hopf") bifurcation occurs when a pair of complex eigenvalues "crosses the imaginary axis" as a parameter is moved (and, in dimensions, bigger than two, the remaining eigenvalues have negative real part), provided that some additional technical conditions hold. (These conditions tend to be satisfied in examples.)

It is very easy to understand the basic idea.

We consider a system:

$$\frac{dx}{dt} = f_\mu(x)$$

in which a parameter "μ" appears.

We assume that the system has dimension two.

Suppose that there are a value μ_0 of this parameter, and a steady state x_0, with the following properties:

• For $\mu < \mu_0$, the linearization at the steady state x_0 is stable, and there is a pair of complex conjugate eigenvalues with negative real part.

• As μ changes from negative to positive, the linearization goes through having a pair of purely imaginary eigenvalues (at $\mu = \mu_0$) to having a pair of complex conjugate eigenvalues with positive real part.

Thus, near x_0, the motion changes from a stable spiral to an unstable spiral as μ crosses μ_0.

If the steady state happens to be a sink even when $\mu = \mu_0$, it must mean that there are nonlinear terms "pushing back" towards x_0 (see the example below).

These terms will still be there for $\mu > \mu_0$, $\mu \approx \mu_0$.

Thus, the spiraling-out trajectories cannot go very far, and a limit cycle is approached.

(Another way to think of this is that, in typical biological problems, trajectories cannot escape to infinity, because of conservation of mass, etc.)

In arbitrary dimensions, the situation is similar. One assumes that all other $n - 2$ eigenvalues have negative real part, for all μ near μ_0.

The $n - 2$ everywhere-negative eigenvalues have the effect of pushing the dynamics towards a two-dimensional surface that looks, near x_0, like the space spanned by the two complex conjugate eigenvectors corresponding to the purely imaginary eigenvalues at $\mu = \mu_0$.

On this surface, the two-dimensional argument that we just gave can be applied.

Let us give more details.

Consider the example that we met earlier:

$$\begin{aligned} dx_1/dt &= \mu x_1 - \omega x_2 + \theta x_1(x_1^2 + x_2^2) \\ dx_2/dt &= \omega x_1 + \mu x_2 + \theta x_2(x_1^2 + x_2^2) \end{aligned}$$

With $\theta = -1$, this is the "supercritical Hopf bifurcation" case in which we go, as already shown, from a globally asymptotically stable equilibrium to a limit cycle as μ crosses from negative to positive (μ_0 is zero).

In contrast, suppose now that $\theta = 1$. The magnitude satisfies the equation $d\rho/dt = \rho(\mu + \rho^2)$.

Hence, one goes again from stable to unstable as μ goes through zero, but now an *unstable* cycle encircles the origin for $\mu < 0$ (so, the origin is not globally attractive).

For $\mu \geq 0$, there is now no cycle that prevents solutions that start near zero from escaping very far. (Once again, in typical biochemical problems, solutions cannot go to infinity. So, for example, a limit cycle of large magnitude might perhaps appear for $\mu > 0$.)

These pictures shows what happens for each fixed value of μ for the supercritical (limit cycle occurs after going from stable to unstable) and subcritical (limit cycle occurs before μ_0) cases, respectively:
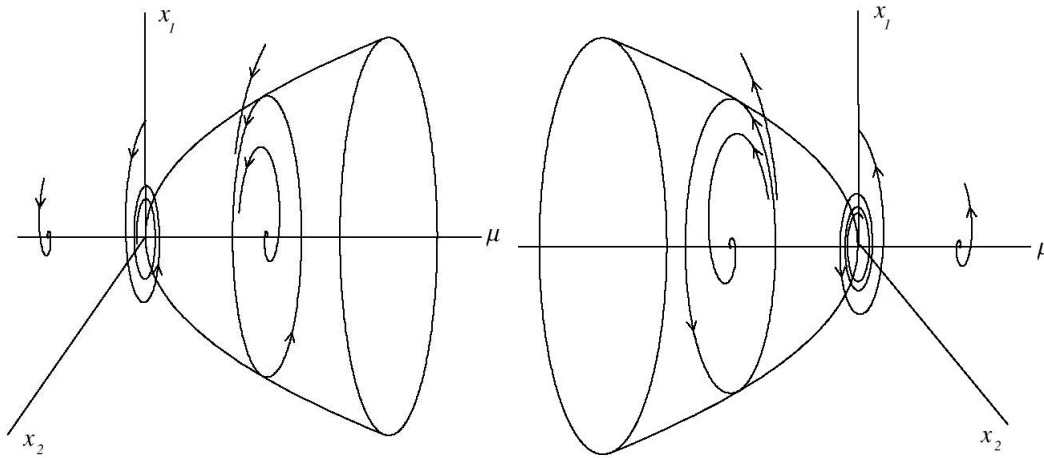
Now suppose given a general system (I will not ask questions in tests about this material; it is merely FYI)[58]:

$$dx/dt = f(x, \mu)$$

in dimension 2, where μ is a scalar parameter and f is assumed smooth. Suppose that for all μ near zero there is a steady-state $\xi(\mu)$, with eigenvalues $\lambda(\mu) = r(\mu) \pm i\omega(\mu)$, with $r(0) = 0$ and $\omega(0) = \omega_0 > 0$, and that $r'(0) \neq 0$ ("eigenvalues cross the imaginary axis with nonzero velocity") and that the quantity α defined below is nonzero. Then, up to a local topological equivalence and time-reparametrization, one can reduce the system to the form given in the previous example, and there is a Hopf bifurcation, supercritical or subcritical depending on $\theta = $ the sign of α.[59] There is no need to perform the transformation, if all we want is to decide if there is a Hopf bifurcation. The general "recipe" is as follows.

Let A be the Jacobian of f evaluated at $\xi_0 = \xi(0)$, $\mu = 0$. and find two complex vectors p, q such that

$$Aq = i\omega_0 q, \quad A^T p = -i\omega_0 p, \quad p \cdot q = 1.$$

Compute the dot product $H(z, \bar{z}) = p \cdot F(\xi_0 + zq + \bar{z}\bar{q}, \mu(0))$ and consider the formal Taylor series:

$$H(z, \bar{z}) = i\omega_0 z + \sum_{j+k \geq 2} \frac{1}{j!k!} g_{jk} z^j \bar{z}^k.$$

---

[58] See e.g. Yu.A. Kuznetsov. Elements of Applied Bifurcation Theory. 2nd ed., Springer-Verlag, New York, 1998

[59] One may interpret the condition on α in terms of a Lyapunov function that guarantees stability at $\mu = 0$, for the supercritical case; see e.g.: Mees , A.I. Dynamics of Feedback Systems, John Wiley & Sons, New York, 1981.

Then $\alpha = \dfrac{1}{2\omega_0^2} \operatorname{Re}\left(ig_{20}g_{11} + \omega_0 g_{21}\right)$.

One may use the following Maple commands, which are copied from "NLDV computer session XI: Using Maple to analyse Andronov-Hopf bifurcation in planar ODEs," by Yu.A. Kuznetsov, Mathematical Institute, Utrecht University, November 16, 1999. They are illustrated by the following chemical model (Brusselator):

$$dx_1/dt = A - (B+1)x_1 + x_1^2 x_2, \;\; dx_2/dt = Bx_1 - x_1^2 x_2$$

where one fixes $A > 0$ and takes B as a bifurcation parameter. The conclusion is that at $B = 1 + A^2$ the system exhibits a supercritical Hopf bifurcation.

```
restart:
with(linalg):
readlib(mtaylor):
readlib(coeftayl):
F[1]:=A-(B+1)*X[1]+X[1]^2*X[2];
F[2]:=B*X[1]-X[1]^2*X[2];
J:=jacobian([F[1],F[2]],[X[1],X[2]]);
K:=transpose(J);
sol:=solve({F[1]=0,F[2]=0},{X[1],X[2]});
assign(sol);
T:=trace(J);
diff(T,B);
sol:=solve({T=0},{B});
assign(sol);
assume(A>0);
omega:=sqrt(det(J));
ev:=eigenvects(J,'radical');
q:=ev[1][3][1];
et:=eigenvects(K,'radical');
P:=et[2][3][1];
s1:=simplify(evalc(conjugate(P[1])*q[1]+conjugate(P[2])*q[2]));
c:=simplify(evalc(1/conjugate(s1)));
p[1]:=simplify(evalc(c*P[1]));
p[2]:=simplify(evalc(c*P[2]));
simplify(evalc(conjugate(p[1])*q[1]+conjugate(p[2])*q[2]));
F[1]:=A-(B+1)*x[1]+x[1]^2*x[2];
F[2]:=B*x[1]-x[1]^2*x[2];
# use z1 for the conjugate of z:
x[1]:=evalc(X[1]+z*q[1]+z1*conjugate(q[1]));
x[2]:=evalc(X[2]+z*q[2]+z1*conjugate(q[2]));
H:=simplify(evalc(conjugate(p[1])*F[1]+conjugate(p[2])*F[2]));
# get Taylor expansion:
g[2,0]:=simplify(2*evalc(coeftayl(H,[z,z1]=[0,0],[2,0])));
g[1,1]:=simplify(evalc(coeftayl(H,[z,z1]=[0,0],[1,1])));
g[2,1]:=simplify(2*evalc(coeftayl(H,[z,z1]=[0,0],[2,1])));
```

```
alpha:=factor(1/(2*omega^2)*Re(I*g[2,0]*g[1,1]+omega*g[2,1]));
evalc(alpha);
# above needed to see that this is a negative number (so supercritical)
```

2.9.4   Combinations of bifurcations

A supercritical Hopf bifurcation is "soft" in that small-amplitude periodic orbits are created. Supercritical bifurcations may give rise to "hard" (big-jump) behavior when embedded in additional fold bifurcations:
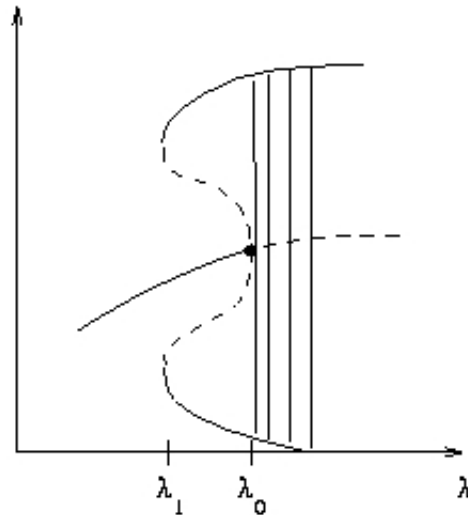
Notice that a "sudden big oscillation" appears.

An example is given by a model of a CSTR[60] as follows:

$$
\begin{aligned}
dy_1/dt &= -y_1 + \mathrm{Da}(1 - y_1)\exp(y_2) \\
dy_2/dt &= -y_2 + B \cdot \mathrm{Da}(1 - y_1)\exp(y_2) - \beta y_2
\end{aligned}
$$

Here,

- y_1, y_2 describe the material and energy balances;

- β is the heat transfer coefficient ($\beta = 3$);

- Da is the Damköhler number (bifurcation parameter: $\lambda :=$Da);

- B is the rise in adiabatic temperature ($B = 16.2$).

The bifurcation diagram is as follows (showing the value of y_1 versus Da:

---

[60]taken from http://www.bifurcation.de/exd2/HTML/exd2.ok.html: A. Uppal, W.H. Ray, A.B. Poore. On the dynamic behavior of continuous stirred tank reactors. Chem. Eng. Sci. 29 (1974) 967-985.
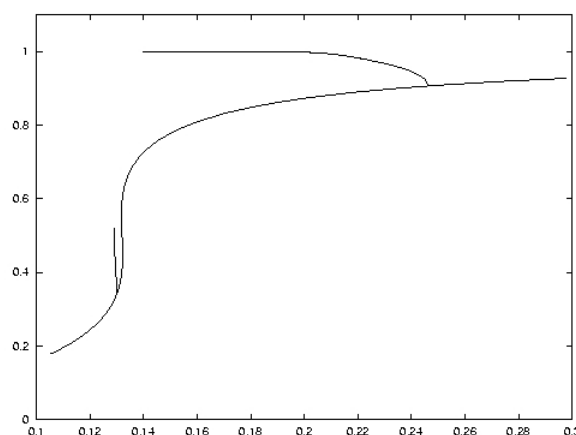
Note that there is first a sub-, and then a super-critical bifurcation. (There is a fold as well, not seen in picture - the first branch continues "backward"). The right one is a supercritical Hopf as the parameter *diminishes*.

A parameter sweep can be used to appreciate the phenomenon: we "sweep" the parameter λ, increasing it very slowly, and simulate the system (for example, by adding an equation $d\lambda/dt = \varepsilon << 1$) With $\varepsilon = 0.001, y_1(0) = 0.1644, y_2(0) = 0.6658$:
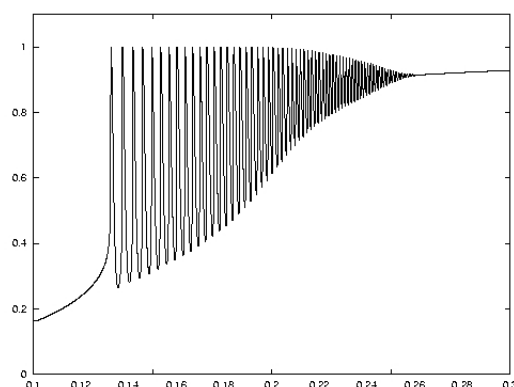
Note the "hard" onset of oscillations (and "soft" end).

Hopf intuition: more dimensions

The Hopf story generalizes to $n > 2$ dimensions. Suppose there are $n - 2$ eigenvalues with negative real part, for μ near μ_0. These $n - 2$ negative eigendirections push the dynamics towards a two-dimensional surface that looks, near x_0, like the space spanned by the two complex conjugate eigenvectors corresponding to the purely imaginary eigenvalues at $\mu = \mu_0$.

On this surface, the two-dimensional argument that we gave can be applied.
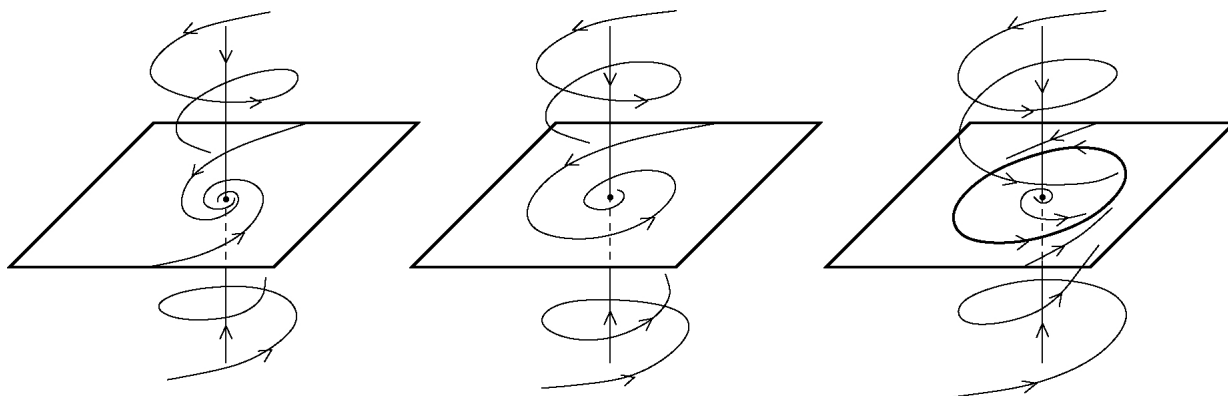
Numerical packages

Numerical packages for bifurcation analysis use continuation methods from a given steady state (and parameter value), testing conditions (singularity of Jacobian, eigenvalues) along the way. As an example, this is typical output using the applet from:

http://techmath.uibk.ac.at/numbau/alex/dynamics/bifurcation/index.html
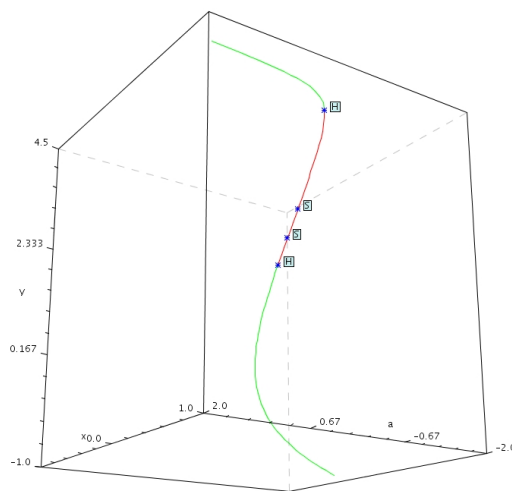
Labeled are points where bifurcations occur.

2.9.5 Cubic Nullclines and Relaxation Oscillations

Let us consider this system, which is exactly as in our version of the van der Pol oscillator, except that, before, we had $\varepsilon = 1$:

$$\frac{dx}{dt} = y + x - \frac{x^3}{3}$$
$$\frac{dy}{dt} = -\varepsilon x$$

We are interested specifically in what happens when ε is positive but small ("$0 < \varepsilon \ll 1$").

Notice that then y changes slowly.

So, we may think of y as a "constant" in so far as its effect on x (the "faster" variable) is concerned. How does $\frac{dx}{dt} = f_a(x) = a + x - \frac{x^3}{3}$ behave?
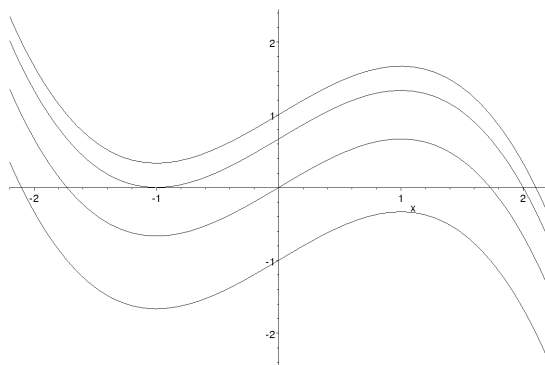
$$f_a(x) = a + x - \frac{x^3}{3} \text{ for } a = -1, 0, \frac{2}{3}, 1$$

$$a = -1$$
$$a = 0$$
$$a = 2/3^-$$
$$a = 2/3^+$$
$$a = 1$$
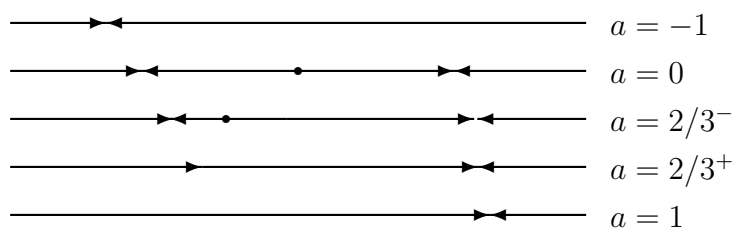
Now let us consider what the solution of the system of differential equations looks like, if starting at a point with $x(0) \ll 0$ and $y(0) \approx -1$.

Since $y(t) \approx -1$ for a long time, x "sees" the equation $dx/dt = f_{-1}(x)$, and therefore $x(t)$ wants to approach a negative "steady state" x_a (approximately at -2)

(If y would be constant, indeed $x(t) \to x_a$.)

However, "a" is not constant, but it is slowly increasing ($y' = -\varepsilon x > 0$).

Thus, the "equilibrium" that x is getting attracted to is slowly moving closer and closer to -1,

until, at exactly $a = 2/3$, the "low" equilibrium dissappears, and there is only the "large" one (around $x = 2$); thus x will quickly converge to that larger value.

Now, however, $x(t)$ is positive, so $y' = -\varepsilon x < 0$, that is, "a" starts *decreasing*.

Repeating this process, one obtains a periodic motion in which slow increases and decreases are interspersed with quick motions.

This is what is often called a *relaxation* (or "hysteresis-driven") oscillation.

Here are computer plot of $x(t)$ for one such solution, together the same solution in phase-plane:
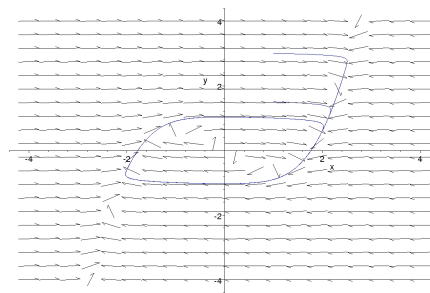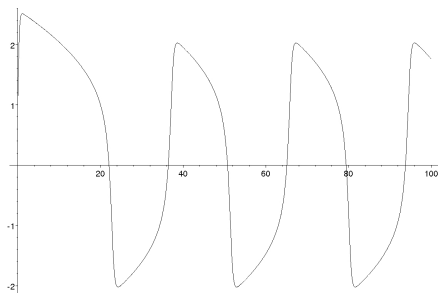
2.9.6 A Qualitative Analysis using Cubic Nullclines

Let us now analyze a somewhat more general situation.

We will assume given a system of this general form:

$$\frac{dx}{dt} = f(x) - y$$
$$\frac{dy}{dt} = \varepsilon \left(g(x) - y \right)$$

where $\varepsilon > 0$. (Soon, we will assume that $\varepsilon \ll 1$, but not yet.)

The x and y nullclines are, respectively: $y = f(x)$ and $y = g(x)$.

It is easy, for these very special equations, to determine the direction of arrows: dy/dt is positive if $y < g(x)$, i.e. under the graph of g, and so forth.

This allows us to draw "SE", etc, arrows as usual:
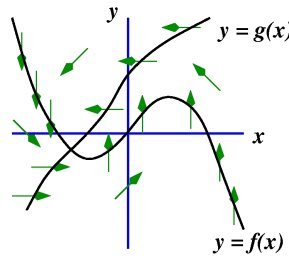
Now let us use the information that ε is small: this means that

   dy/dt is always very small compared to dx/dt, i.e., the arrows are (almost) horizontal,

except very close to the graph of $y=f(x)$, where both are small (exactly vertical, when $y=f(x)$):
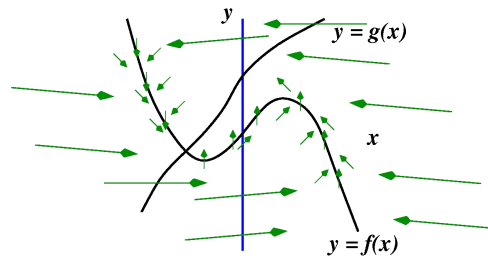
Now, suppose that the nullclines look exactly as in these pictures, so that $f' < 0$ and $g' > 0$ at the steady state.

The Jacobian of $\begin{pmatrix} f(x) - y \\ \varepsilon(g(x) - y) \end{pmatrix}$ is

$$\begin{pmatrix} f'(x_0) & -1 \\ \varepsilon g'(x_0) & -\varepsilon \end{pmatrix}$$

and therefore (remember that $f'(x_0) < 0$) the trace is negative, and the determinant is positive (because $g'(x_0) > 0$), and the steady state is a sink (stable).

Thus, we expect trajectories to look like this:
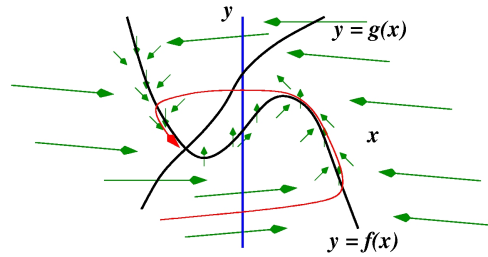
Observe that a "large enough" perturbation from the steady state leads to a large excursion (the trajectory is carried very quicky to the other side) before the trajectory can return.
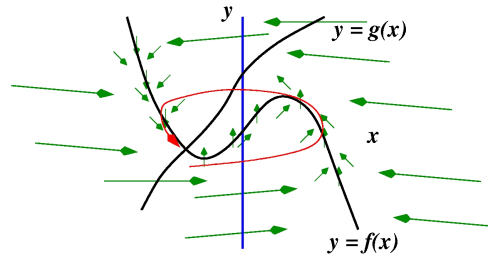
In contrast, a small perturbation does not result in such excursions, since the steady state is stable. Zooming-in:

This type of behavior is called *excitability*: low enough disturbances have no effect, but when over a threshold, a large reaction occurs.

In contrast, suppose that the nullcline $y = g(x)$ intersects the nullcline $y = f(x)$ on the increasing part of the latter ($f' > 0$).

Then, the steady state is *un*stable, for small ε, since the trace is $f'(x_0) - \varepsilon \approx f'(x_0) > 0$. In fact, it is a repelling state, because the determinant of the Jacobian equals $\varepsilon(g'(x_0) - f'(x_0)) > 0$ (notice in the figure that $g' > f'$ at the intersection of the plots).

In any case, it is clear by "following directions" that we obtain a relaxation oscillation, instead of an excitable system, in this case:
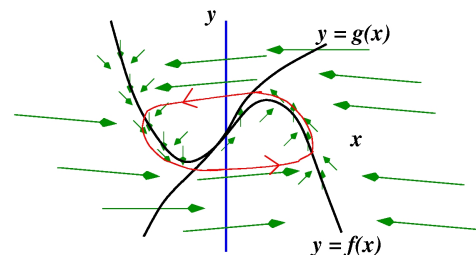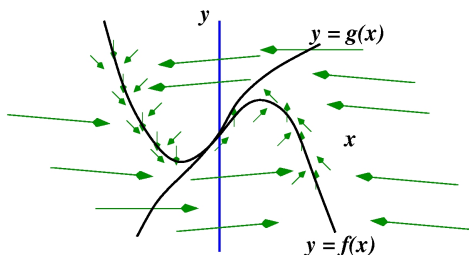
2.9.7 Neurons

Neurons are nerve cells; there are about 100 billion (10^{11}) in the human brain.

Neurons may be short (1mm) or very long (1m from the spinal cord to foot muscles).

Each neuron is a complex information processing device, whose inputs are neurotransmitters (electrically charged chemicals) which accumulate at the *dendrites*.

Neurons receive signals from other neurons (from as many as 150,000, in the cerebral cortex, the center of cognition) connected to it at *synapses*.

When the net voltage received by a neuron is higher than a certain threshold (about 1/10 of a volt), the neuron "fires" an *action potential*, which is an electrical signal that travels down the *axon*, sort of an "output wire" of the neuron. Signals can travel at up to 100m/s; the higher speeds are achieved when the axon is covered in a fatty insulation (myelin).

At the ends of axons, neurotransmitters are released into the dendrites of other neurons.

Information processing and computation arise from these networks of neurons.
The strength of synaptic connections is one way to "program" networks; memory (in part) consists of finely tuning these strengths.
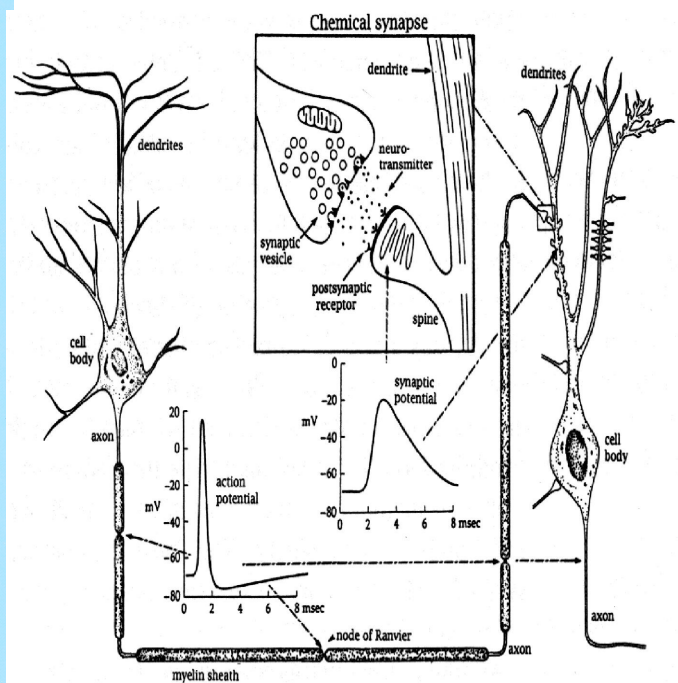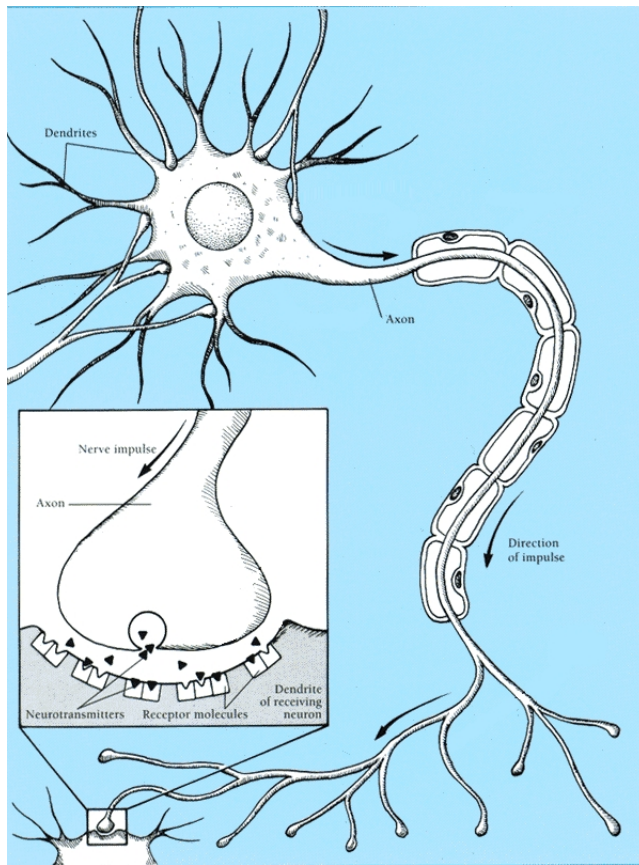
The mechanism for action potential generation is well understood. A mathematical model given in: Hodgkin, A.L. and Huxley, A.F., "A Quantitative Description of Membrane Current and its Application to Conduction and Excitation in Nerve", Journal of Physiology 117 (1952): 500-544 won the authors a Nobel Prize (in 1963), and is still one of the most successful examples of mathematical modeling in biology. Let us sketch it next.

2.9.8 Action Potential Generation

The basic premise is that currents are due to Na and K ion pathways. Normally, there is more K^+ inside than outside the cell, and the opposite holds for Na^+. Diffusion through channels works against this imbalance, which is maintained by active pumps (which account for about 2/3 of the cell's energy

consumption!). These pumps act against a steep gradient, exchanging 3 Na^+ ions out for each 2 K^+ that are allowed in. An overall potential difference of about 70mV is maintained (negative inside the cell) when the cell is "at rest".

A neuron can be stimulated by external signals (touch, taste, etc., sensors), or by an appropriate weighted sum of inhibitory and excitatory inputs from other neurons through dendrites (or, in the Hodgkin-Huxley and usual lab experiments, artificially with electrodes).

The key components are *voltage-gated ion channels*[61]:
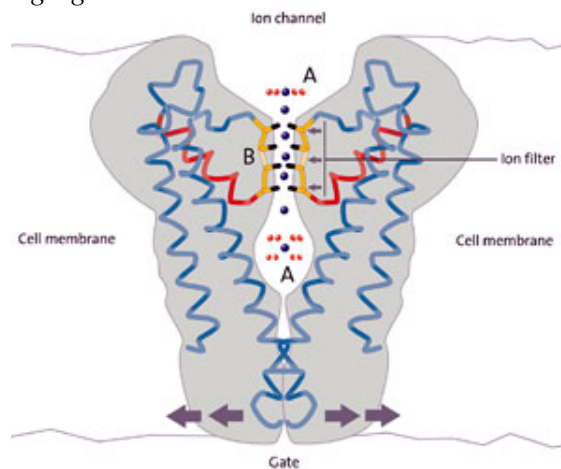
A large enough potential change triggers a nerve impulse (action potential or "spike"), starting from the axon hillock (start of axon) as follows:

(1) voltage-gated Na^+ channels open (think of a "gate" opening); these let sodium ions in, so the inside of the cell becomes more positive, and, through a feedback effect, even more gates open;

(2) when the voltage difference is $\approx +50$mV, voltage-gated K^+ channels open and quickly let potassium out;

(3) the Na^+ channels close;

(4) the K^+ channels close, so we are back to resting potential.

The Na^+ channels cannot open again for some minimum time, giving the cell a *refractory period*.

Some pictures follow, illustrating the same important process in slightly different ways.

---
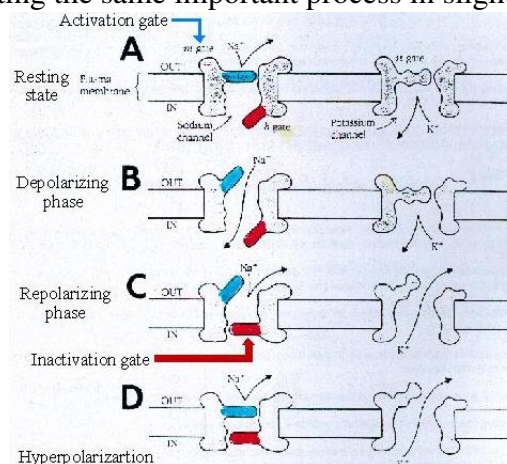
[61] Illustration from http://fig.cox.miami.edu/ cmallery/150/memb/ion_channel1_sml.jpg

(http://jimswan.com/237/channels/channel_graphics.htm)

(http://www.cellscience.com/reviews3/spikes.jpg)
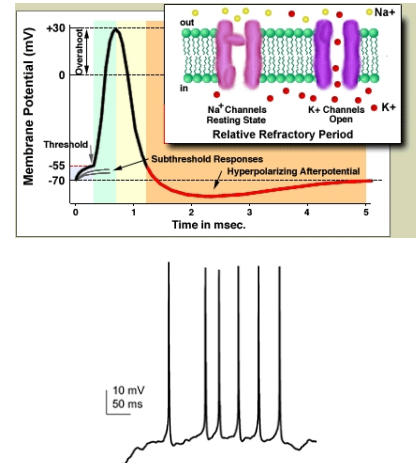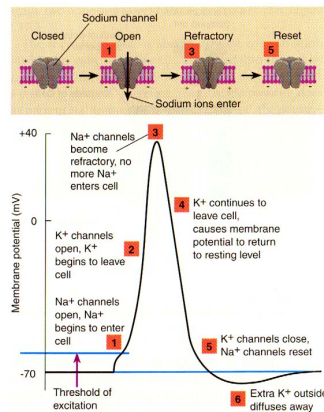
This activity, locally in the axon, affects neighboring areas, which then go through the same process, a chain-reaction along the axon. Because of the refractory period, the signal "cannot go back", and a direction of travel for the signal is well-defined.

(Copyright 1997, Carlos Finlay and Michael R. Markham).

These diagrams are from http://www.biologymad.com/NervousSystem/nerveimpulses.htm:
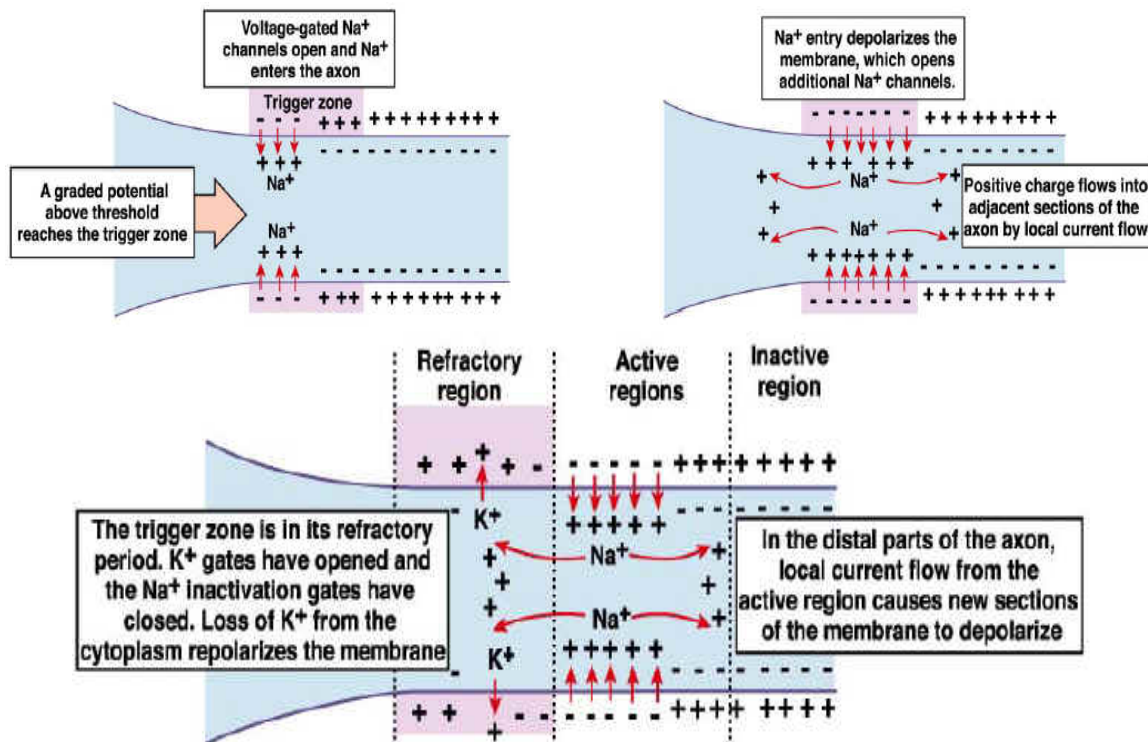
It is important to realize that the action potential is only generated if the stimulus is large enough. It is an "all or (almost) nothing" response. An advantage is that the signal travels along the axon without decay - it is regenerated along the way. The "binary" (digital) character of the signal makes it very robust to noise.

There is another aspect that is remarkable, too: a continuous stimulus of high intensity will result in a higher frequency of spiking. Amplitude modulation (as in AM radio) gets transformed into frequency modulation (as in FM radio, which is far more robust to noise).

2.9.9 Model

The basic HH model is for a small segment of the axon. Their model was done originally for the giant axon of the squid (large enough to stick electrodes into, with the technology available at the time), but similar models have been validated for other neurons.

(Typical simulations put together perhaps thousands of such basic compartments, or alternatively set up a partial differential equation, with a spatial variable to represent the length of the axon.)

The model has four variables: the potential difference $v(t)$ between the inside and outside of the neuron, and the activity of each of the three types of gates (two types of gates for sodium and one for potassium). These activities may be thought of as relative fractions ("concentrations") of open channels, or probabilities of channels being open. There is also a term I for the external current being applied.

$$
\begin{aligned}
C dv/dt &= -g_K(t)(v-v_K) - g_{Na}(t)(v-v_{Na}) - \bar{g}_L(v-v_L) + I \\
\tau_m(v) dm/dt &= m_\infty(v) - m \\
\tau_n(v) dn/dt &= n_\infty(v) - n \\
\tau_h(v) dh/dt &= h_\infty(v) - h \\
g_K(t) &= \bar{g}_K\, n(t)^4 \\
g_{Na}(t) &= \bar{g}_{Na}\, m(t)^3\, h(t)
\end{aligned}
$$

The equation for v comes from a capacitor model of membranes as charge storage elements. The first three terms in the right correspond to the currents flowing through the Na and K gates (plus an additional "L" that accounts for all other gates and channels, not voltage-dependent).

The currents are proportional to the difference between the actual voltage and the "Nernst potentials" for each of the species (the potential that would result in balance between electrical and chemical imbalances), multiplied by "conductances" g that represent how open the channels are.

The conductances, in turn, are proportional to certain powers of the open probabilities of the different gates. (The powers were fit to data, but can be justified in terms of cooperativity effects.)

The open probabilities, in turn, as well as the time-constants (τ's) depend on the current net voltage difference $v(t)$. H&H found the following formulas by fitting to data. Let us write:

$$
\frac{1}{\tau_m(v)}\left(m_\infty(v) - m\right) = \alpha_m(v)(1-m) - \beta_m(v)m
$$

(so that $dm/dt = \alpha_m(v)(1-m) - \beta_m(v)m$), and similarly for n, h. In terms of the α's and β's, H&H's formulas are as follows:

$$
\alpha_m(v) = 0.1\frac{25 - v}{\exp\left(\frac{25-v}{10}\right) - 1}, \quad \beta_m(v) = 4\exp\left(\frac{-v}{18}\right), \quad \alpha_h(v) = 0.07\exp\left(\frac{-v}{20}\right),
$$

$$
\beta_h(v) = \frac{1}{\exp\left(\frac{30-v}{10}\right) + 1}, \quad \alpha_n(v) = 0.01\frac{10 - v}{\exp\left(\frac{10-v}{10}\right) - 1}, \quad \beta_n(v) = 0.125\exp\left(\frac{-v}{80}\right)
$$

where the constants are $\bar{g}_K = 36$, $\bar{g}_{Na} = 120$, $\bar{g}_L = 0.3$ $v_{Na} = 115$ $v_K = -12$, and $v_L = 10.6$.

The way in which H&H did this fit is, to a large extent, the best part of the story. Basically, they performed a "voltage clamp" experiment, by inserting an electrode into the axon, thus permitting a plot of current against voltage, and deducing conductances for each channel. (They needed to isolate the effects of the different channels; the experiments are quite involved, and we don't have time to go over them in this course.)

For an idea of how good the fits are, look at these plots of experimental $g_K(V)(t)$ and $g_{Na}(V)(t)$, for different clamped V's (circles) compared to the model predictions (solid curves).
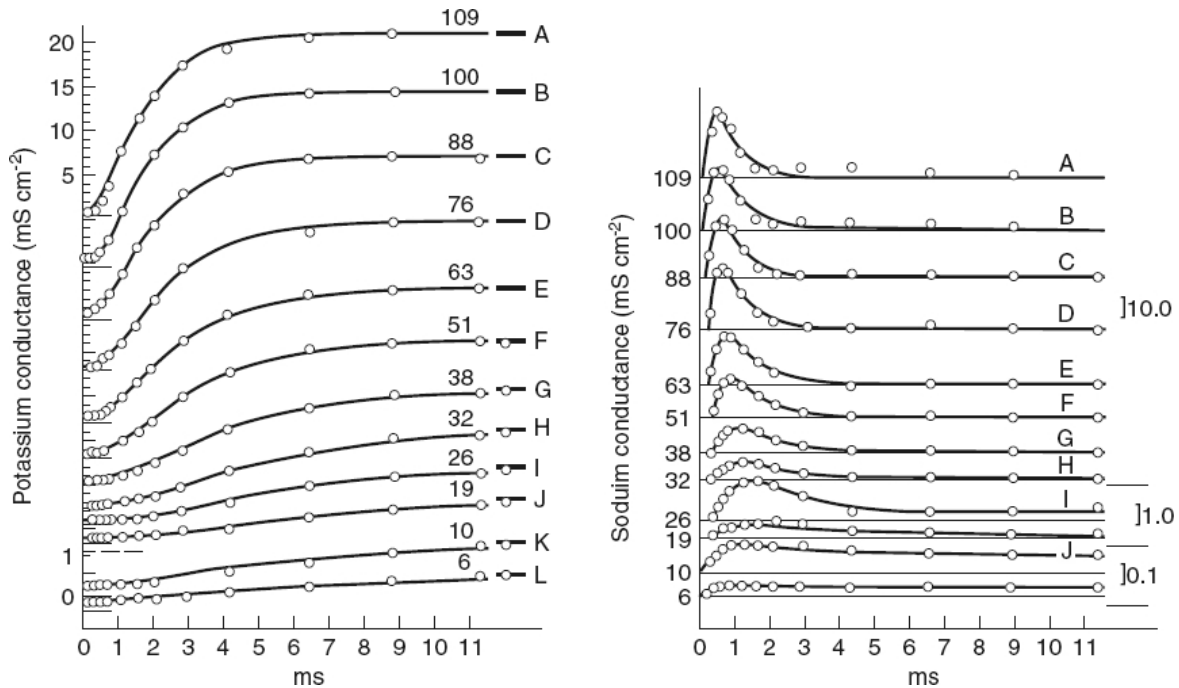
Simulations of the system show frequency encoding of amplitude.
We show here the responses to constant currents of 0.05 (3 spikes in the shown time-interval), 0.1 (4), 0.15 (5) mA:
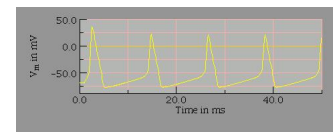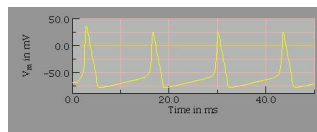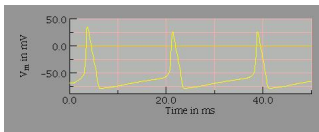
Here are the plots of n, m, h in response to a stimulus at $t = 5$ of duration 1sec, with current=0.1:
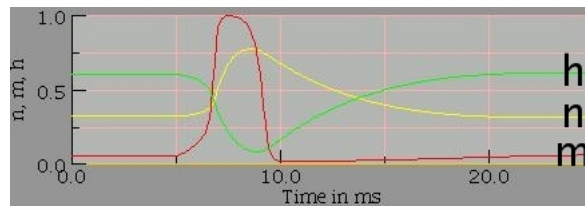
(color code: yellow=n, red=m, green=h)

Observe how m moves faster in response to stimulus.

It is an important feature of the model that $\tau_m \ll \tau_n$ and $\ll \tau_h$. This allows a time-scale separation analysis (due to FitzHugh): for short enough intervals, one may assume that $n(t) \equiv n_0$ and $h \equiv h_0$,

so we obtain just two equations:

$$Cdv/dt = -\bar{g}_K n_0^4(v-v_K) - \bar{g}_{Na}m^3h_0(v-v_{Na}) - \bar{g}_L(v-v_L)$$
$$\tau_m(v)dm/dt = m_\infty(v) - m.$$

The phase-plane shows bistability (dash-dot curve is separatrix, dashed curve is nullcline $dv/dt = 0$, dotted curve is nullcline $dm/dt = 0$; two solutions are shown with a solid curve)[62]:
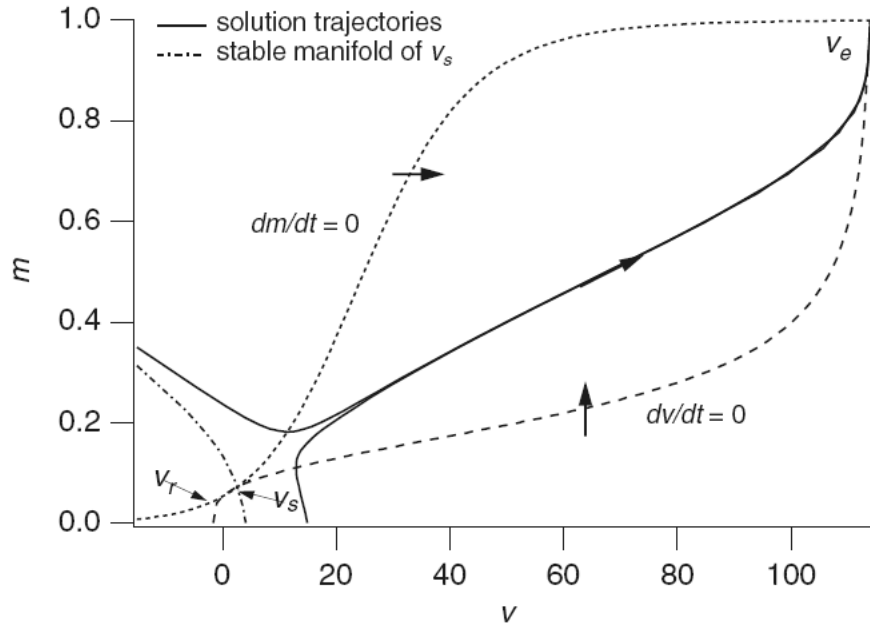
There are two stable steady states: v_r ("resting") and v_e ("excited"), as well as a saddle v_s. Depending on where the initial voltage (set by a transient current I) is relative to a separatrix, as $t \to \infty$ trajectories either converge to the "excited" state, or stay near the resting one.

(Of course, h, n are not really constant, so the analysis must be complemented with consideration of small changes in h, n. We do not provide details here.)

An alternative view, on a longer time scale, is also possible. FitzHugh observed (and you will, too, in an assigned project; see also the graph shown earlier) that : $h(t) + n(t) \approx 0.8$, constant during an action potential. (Notice the approximate symmetry of h, n in plots.) This allows one to eliminate h from the equations. Also, assuming that $\tau_m \ll 1$ (because we are looking at a longer time scale), we may replace $m(t)$ by its quasi-steady state value $m_\infty(v)$. We end up with a new two-dimensional system:

$$Cdv/dt = -\bar{g}_K n^4(v - v_K) - \bar{g}_{Na}m_\infty(v)^3(0.8 - n)(v - v_{Na}) - \bar{g}_L(v - v_L)$$
$$\tau_n(v)dn/dt = n_\infty(v) - n$$

which has these nullclines (dots for $dn/dt=0$, dashes for $dv/dt=0$) and phase plane behavior:

---

[62]next two plots borrowed from Keener & Sneyd textbook
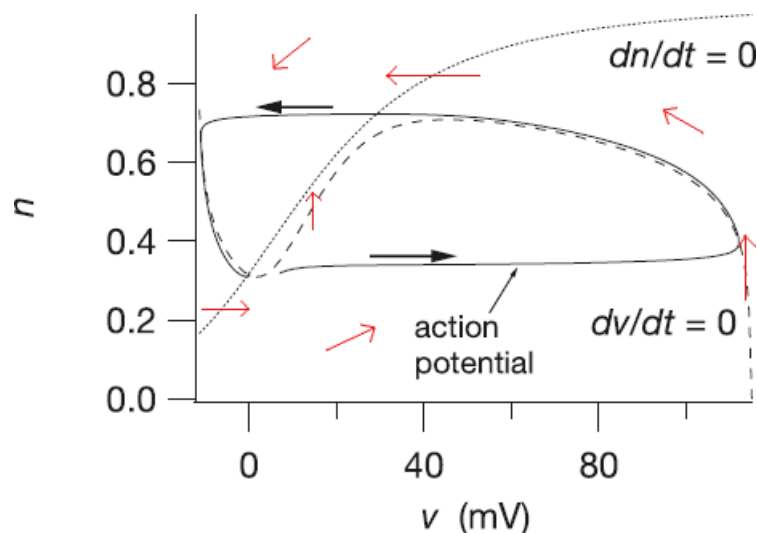
We have fast behaviors on the horizontal direction (n=constant), leading to v approaching nullclines fast, with a slow drift on n that then produces, as we saw earlier when studying a somewhat simpler model of excitable behavior, a "spike" of activity.

Note that if the nullclines are perturbed so that they now intersect in the middle part of the "cubic-looking" curve (for v, this would be achieved by considering the external current I as a constant), then a relaxation oscillator will result. Moreover, if the perturbation is larger, so that the intersection is away from the "elbows", the velocity of the trajectories should be higher (because trajectories do not slow-down near the steady state). This explains "frequency modulation" as well.

Much of the qualitative theory of relaxation oscillations and excitable systems originated in the analysis of this example and its mathematical simplifications.

The following website:

http://www.scholarpedia.org/article/FitzHugh-Nagumo_model

has excellent animated-gifs showing the processes.

2.10 Problems for ODE chapter

Problems ODE1: Population growth

1. Show that the solution of the logistic equation, given by $N(t) = \frac{N_0 B}{N_0 + (B - N_0)e^{-rt}}$, has the following properties:

   (a) $N \to B$ as $t \to \infty$.

   For the next parts, you may use the fact that, if N is a solution of $\frac{dN}{dt}(t) = f(N(t))$, then $\frac{d^2 N}{dt^2}(t) = f'(N(t))f(N(t))$ (as remarked in the next problem).

   (b) The graph of $N(t)$ is concave up when $N < \frac{B}{2}$.

   (c) The graph of $N(t)$ is concave down for $\frac{B}{2} < N < B$.

   (d) The graph of $N(t)$ is concave up when $N > B$.

2. Sometimes, with large populations competing for limited resources, this competition might result in struggles that prevent reproduction, so that $\frac{dN}{dt}(t)$ is actually smaller if $N(t)$ is larger. (Another example is tumor cells that have less access to blood vessels when the volume of the tumor is large.) This means that $\frac{d^2 N}{dt^2} < 0$ if N is large, and is sometimes called "decelerating growth for large populations" ("decreasing growth" would be a more accurate term!), though other authors may use that term with a slightly different meaning.[63]

   Observe that, if N is a solution of $\frac{dN}{dt}(t) = f(N(t))$, then $\frac{d^2 N}{dt^2}(t) = f'(N(t))f(N(t))$.

   Thus, in order to test decelerating growth, we need to test $f'(N)f(N) < 0$ when N is large.

   Let us may write single population models as $\frac{dN}{dt} = K(N)N$, where we think of $K(N)$ as a "per capita" growth rate. In the simplest exponential growth model, $K(N)$ is constant, and in the logistic model, $K(N)$ is a linear function. But many other choices are possible as well. For single-species populations, consider the following per-capita growth rates $K(N)$, and answer for which of them we have decelerating growth:

   (a) $K(N) = \frac{\beta}{1+N}, \beta > 0$.

   (b) $K(N) = \beta - N, \beta > 0$.

   (c) $K(N) = N - e^{\alpha N}, \alpha > 0$.

   (d) $K(N) = \log N$.

   (e) Which of the above growth rates would result in a stable population size?

   (f) Think of examples in biology where these different growth rates may arise.

3. The "Allee effect" in biology (named after American zoologist and ecologist Warder Clyde Allee) is that in which there is a positive correlation between population density and the per capita population growth rate $K(N)$ in very small populations. In Allee-effect models, the function $K(N)$ has a strict maximum at some intermediate value α of density, on the interval $[0, B]$:

$$dK/dN > 0 \text{ for } 0 < N < \alpha, \quad dK/dN < 0 \text{ for } \alpha < N < B$$

---

[63]For example, search in Google books inside the book "Cancer of the breast" by William L. Donegan and John Stricklin Spratt, or the paper "Decelerating growth and human breast cancer" by John A. Spratt, D. von Fournier, John S. Spratt, and Ernst E. Weber, Cancer, Volume 71, pages 2013-2019, March 1993.

(and $K(B) = 0$). Such a model applies when there is a carrying capacity, limiting growth at large densities, as in the logistic model, but also growth is impeded at low densities (for example, because of the difficulty in finding mates at low densities). The sign of $K(0)$ may be positive or negative.

(a) Sketch a plot of a hypothetical function $K(N)$ that satisfies these properties, assuming that $K(0) < 0$. (That is, all we are asking is that $K(0) < 0$, $K(B) = 0$, and there is a maximum at some intermediate point between 0 and B.)

(b) Give an explicit example of a parabola $K(N)$ as in (a).

(c) Sketch a plot of $K(N)N$ for your function K in the previous part.

(d) What are the stable points of $dN/dt = K(N)N$, for such a function $K(N)$?

(e) Repeat a-d with a function that satisfies $K(0) > 0$.

Remarks on this example: If we ask that $K(0) < 0$, then we have something called a "strong" Allee effect: in this case, the growth rate per capita is actually negative for small populations. If instead $K(B) \geq 0$, then one talks about the "weak" Allee effect. The corresponding plots for dN/dt, which is the product $K(N)N$ and not just $K(N)$, are as follows:
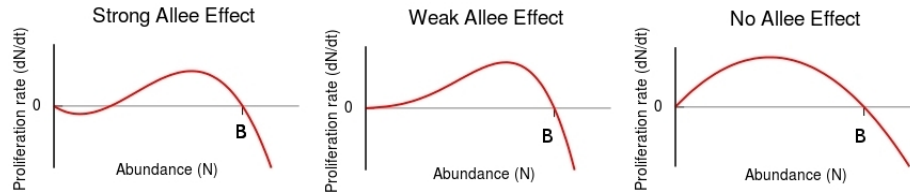
The plot shown below[64] provides experimental evidence of a strong Allee effect, showing the relationship between per capita growth rate and population density for crown-of-thorns starfish (Acanthaster planci):
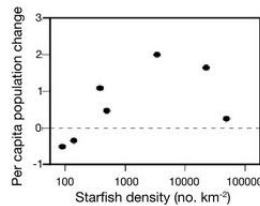
4. The continuous-time Beverton-Holt population growth model in fisheries is:

$$\frac{dN}{dt} = \frac{rN}{\alpha + N}$$

(Beverton and Holt, 1957; Keshet's book).

(a) Find k_1 and k_2 so that, with $N^* = k_1 N$ and $t^* = k_2 t$, we obtain the following "dimensionless" form of the Beverton-Holt model:

$$\frac{dN^*}{dt^*} = \frac{N^*}{1 + N^*} \quad .$$

---

[64]This is quoted from Drake, JM. and Kramer, AM. (2011) Allee Effects. Nature Education Knowledge 3(10):2 (http://www.nature.com/scitable/knowledge/library/allee-effects-19699394), where it is attributed to: Dulvy NK, Freckleton RP, and Polunin NVC (2004) Coral reef cascades and the indirect effects of predator removal by exploitation. Ecol Lett 7:410416.

(If you like, change the statement to "find $\hat{t}$ and $\hat{N}$ such that, with $\hat{N}N^* = N$ and $\hat{t}t^* = t$, ..." – it is good to get used to other notations, though.)

(b) Analyze the behavior of the solutions to this equation.

5. Another population growth model used in fisheries is the continuous Ricker model:

$$\frac{dN}{dt} = rNe^{-\beta N}$$

(Ricker, 1954; Keshet's book). (a) Find k_1 and k_2 so that, with $N^* = k_1 N$ and $t^* = k_2 t$, we obtain the following "dimensionless" form of the Ricker model:

$$\frac{dN^*}{dt^*} = N^* e^{-N^*} \quad .$$

(b) Analyze the behavior of the solutions to this equation.

Problems ODE2: Interacting population problems

When species interact, the population dynamics of each species is affected. (See Section 2.1.11 on "Signs of interactions among variables".) There are three main types of interactions:

- If the growth rate of one of the populations is decreased, and the other increased, the populations are in a *predator-prey* relationship.

- If the growth rate of each population is decreased, then they are in a *competition* relationship.

- If each population's growth rate is enhanced, then they are in a *mutualism* relationship.

1. (**Predator-Prey Models**) The classical version of the *Lotka-Volterra predator-prey system* is as follows:

$$
\begin{aligned}
\frac{dN}{dt} &= N(a - bP) \\
\frac{dP}{dt} &= P(cN - d).
\end{aligned}
$$

Here, $N(t)$ is the prey population and $P(t)$ is the predator population at time t, and a, b, c and d are positive constants.

(a) Briefly explain the ecological assumptions in the model, i.e. interpret each term in the equations.

(b) Find a rescaling of variables that makes the model have just one parameter:

$$
\begin{aligned}
\frac{dN}{dt} &= N(1 - P) \\
\frac{dP}{dt} &= \alpha P(N - 1)
\end{aligned}
$$

(c) Determine the steady states and their stability.

(d)(i) Show that the solution of the differential equation with $d = a$ and initial conditions $N(0) = 2$ and $P(0) = \frac{1}{2}$ satisfies:

$$
N(t) + P(t) - \log\left(N(t)P(t)\right) = \frac{5}{2} \tag{$*$}
$$

for all $t \geq 0$. (Hint: show that the derivative of $N(t) + P(t) - \log\left(N(t)P(t)\right)$ is zero.)

(ii) Plot the solutions of $N(t) + P(t) - \log\left(N(t)P(t)\right) = \frac{5}{2}$. Are the solutions periodic?

(iii) Give a general formula like $(*)$ for arbitrary a, b, c, d, and any initial conditions.

2. (Predator-prey model with limited growth.) Here is a modified version of the classical version of the predator-prey model, in which there is logistic growth for the prey, with carrying capacity K, as follows:

$$
\begin{aligned}
\frac{dN}{dt} &= \alpha\left(1 - \frac{N}{K}\right)N - \gamma PN = \alpha N - \beta N^2 - \gamma PN \\
\frac{dP}{dt} &= -\delta P + \epsilon PN.
\end{aligned}
$$

We have that $\beta = \alpha/K$ and α now represents the inherent per capita net growth rate for the prey in the absence of predators. As in the previous predator-prey model, predators eat only prey, and, if there were no prey, they would die at a constant per capita rate.

We assume that $\alpha/\beta > \delta/\varepsilon$.

(a) Draw the nullclines of the system, assuming that $\alpha = 2$ and $\beta = \delta = \varepsilon = \gamma = 1$. Show the directions of movement "South or North" and "East or West" on the N and P nullclines respectively. Then conclude, and show also in the diagram, the directions of movement ("Northeast," etc.) on each region (connected component) of the first quadrant delimited by the nullclines.

(b) Show that the steady states for the model are $(0,0)$, $\left(\frac{\alpha}{\beta}, 0\right)$, and

$$\left(\frac{\delta}{\epsilon}, \frac{\alpha}{\gamma} - \frac{\beta\delta}{\gamma\epsilon}\right).$$

(c) Form the Jacobian matrix at $\left(\frac{\delta}{\epsilon}, \frac{\alpha}{\gamma} - \frac{\beta\delta}{\gamma\epsilon}\right)$. Show that $\left(\frac{\delta}{\epsilon}, \frac{\alpha}{\gamma} - \frac{\beta\delta}{\gamma\epsilon}\right)$ is a stable steady state.

(d) For the special case $\alpha = 2$ and $\beta = \delta = \varepsilon = \gamma = 1$, determine the stability and sketch the phase plane near the equilibrum $(2,0)$. (If real eigenvalues, find eigenvalues and eigenvectors for the linearization at that point. If complex, determine if clockwise or counterclockwise spiral.)

(e) For the special case $\alpha = 2$ and $\beta = \delta = \varepsilon = \gamma = 1$, determine the stability and sketch the phase plane near the equilibrum $(1,1)$. (If real eigenvalues, find eigenvalues and eigenvectors for the linearization at that point. If complex, determine if clockwise or counterclockwise spiral.)

3. (Another revised predator-prey model.) This example illustrates how the type of equilibrium may change because of changes in the values of the model parameters.

As in the previous model, we have predators P and prey N. The prey growth follows the logistic equation as in the predator-prey model with limited growth, but we change the growth assumption regarding predators : instead of being proportional to NP, it is proportional to $\frac{NP}{1+N}$, a Michaelis-Menten rate:

$$\frac{dN}{dt} = \alpha N - \beta N^2 - \gamma \frac{PN}{1+N}$$

We also change the assumptions on the predator growth rate:

$$\frac{dP}{dt} = \delta P \left(1 - \frac{\epsilon P}{N}\right)$$

reflecting a carrying capacity proportional to the prey population. Consider this system, with the parameters $\alpha = \frac{2}{3}$, $\beta = \frac{1}{6}$, $\gamma = 1$, and $\epsilon = 1$ and answer the following questions.

(a) Draw the nullclines of the system. Show the directions of movement "South or North" and "East or West" on the N and P nullclines respectively. Then conclude, and show also in the diagram, the directions of movement ("Northeast," etc.) on each region (connected component) of the first quadrant delimited by the nullclines.

(b) Show that the point $(1,1)$ is a steady state. Under what condition is it a stable, and under what condition is it an unstable, steady state?

4. (**Competition Models**) We now consider the basic two-species Lotka-Volterra *competition* model with each species N_1 and N_2 having logistic growth in the absence of the other:

$$\frac{dN_1}{dt} = r_1 N_1 \left[1 - \frac{N_1}{k_1} - b_{12}\frac{N_2}{k_1}\right]$$

$$\frac{dN_2}{dt} = r_2 N_2 \left[1 - \frac{N_2}{k_2} - b_{21}\frac{N_1}{k_2}\right]$$

where r_1, r_2, k_1, k_2, b_{12} and b_{21} are all positive constants, the r's are the linear birth rates at low densities, and the k's are the carrying capacities. The constants b_{12} and b_{21} measure the competitive effect of N_2 on N_1 and N_1 on N_2 respectively.

(a) Find a rescaling of variables that makes the model have just three parameters:

$$\frac{dN_1}{dt} = N_1(1 - N_1 - a_{12}N_2)$$

$$\frac{dN_2}{dt} = \alpha N_2(1 - N_2 - a_{21}N_1)$$

For the remainder of this problem we assume that the system is already in this reduced form.

(b) Suppose in this part that $a_{12}a_{21} = 1$.

(i) Show that there are three steady states: $(0,0)$, $(1,0)$, and $(0,1)$. and that $(0,0)$ is unstable.

(ii) Show that $(1,0)$ is stable if $a_{21} > 1$ and unstable if $a_{21} < 1$.

(iii) Under what conditions can you guarantee that $(0,1)$ is stable or unstable?

(c) Suppose in this part that $a_{12} = a_{21} = \frac{1}{2}$.

(i) Show that there are four steady states, and that one of these has both $N_1 > 0$ and $N_2 > 0$ and is stable (a *coexistence* state).

(ii) Draw the nullclines of the system, assuming $a_{12} = a_{21} = \frac{1}{2}$. Show the directions of movement "South or North" and "East or West" on the N_1 and N_2 nullclines respectively. Then conclude, and show also in the diagram, the directions of movement ("Northeast," etc.) on each region (connected component) of the first quadrant delimited by the nullclines.

(d) Suppose in this part that $a_{12} = a_{21} = 2$.

(i) Show that there are four steady states, and one of them has both $N_1 > 0$ and $N_2 > 0$ and is a saddle.

(ii) Draw the nullclines of the system, assuming $a_{12} = a_{21} = 2$. Show the directions of movement "South or North" and "East or West" on the N_1 and N_2 nullclines respectively. Then conclude, and show also in the diagram, the directions of movement ("Northeast," etc.) on each region (connected component) of the first quadrant delimited by the nullclines.

5. (**Mutualism or Symbiosis Models**) There are many examples that show the interaction of two or more species has advantages for all. Mutualism often plays the crucial role in promoting and even maintaining such species: plant and seed dispersal is one example. The simplest mutualism model analogous to the classical Lotka-Volterra predator-prey one is as follows:

$$\frac{dN_1}{dt} = r_1 N_1 + a_1 N_1 N_2$$

$$\frac{dN_2}{dt} = r_2 N_2 + a_2 N_2 N_1$$

where r_1, r_2, a_1 and a_2 are all positive constants.

Determine the steady states and their stabilities.

6. (a) Determine the kind of interactive behavior between two species with populations N_1 and N_2 that is implied by the following model:

$$\frac{dN_1}{dt} = r_1 N_1 \left(1 - \frac{N_1}{k_1 + b_{12} N_2} \right)$$

$$\frac{dN_2}{dt} = r_2 N_2 \left(1 - \frac{N_2}{k_2 + b_{21} N_1} \right)$$

where r_1, r_2, k_1, k_2, b_{12} and b_{21} are all positive constants.

(b) Find a rescaling of variables that makes the model have just three parameters:

$$\frac{dN_1}{dt} = N_1 \left(1 - \alpha \frac{N_1}{1 + N_2} \right)$$

$$\frac{dN_2}{dt} = \gamma N_2 \left(1 - \beta \frac{N_2}{1 + N_1} \right)$$

For the remainder of this problem we assume that the system is already in this reduced form.

(c) Determine the steady states of the system and their stabilities.

(d) Draw the nullclines of the system, assuming the special values $r_1 = r_2 = k_1 = k_2 = 1$ and $b_1 = b_2 = 1/2$. Show the directions of movement "South or North" and "East or West" on the N_1 and N_2 nullclines respectively. Then conclude, and show also in the diagram, the directions of movement ("Northeast," etc.) on each region (connected component) of the first quadrant delimited by the nullclines.

7. (a) Determine the kind of interactive behavior between two species with populations N_1 and N_2 that is implied by the following model:

$$\frac{dN_1}{dt} = r N_1 \left(1 - \frac{N_1}{k} \right) - a N_1 N_2 \left(1 - \exp(-b N_1) \right)$$

$$\frac{dN_2}{dt} = -d N_2 + N_2 e \left(1 - \exp(-b N_1) \right)$$

where r, a, b, d, e and k are positive constants.

(b) Find a rescaling of variables that makes the model have just three parameters:

$$\frac{dN_1}{dt} = N_1 (1 - N_1) - N_1 N_2 \left(1 - e^{-\beta N_1} \right)$$

$$\frac{dN_2}{dt} = -\gamma N_2 + \alpha N_2 \left(1 - e^{-\beta N_1} \right)$$

For the remainder of this problem we assume that the system is already in this reduced form.

(c) Determine the steady states of the system and their stabilities.

(d) Draw the nullclines of the system. Show the directions of movement "South or North" and "East or West" on the N_1 and N_2 nullclines respectively. Then conclude, and show also in the diagram, the directions of movement ("Northeast," etc.) on each region (connected component) of the first quadrant delimited by the nullclines.

8. (a) Determine the kind of interactive behavior between two species with populations N_1 and N_2 that is implied by the following model:

$$\frac{dN_1}{dt} = N_1(a - d(N_1 + N_2) - b)$$

$$\frac{dN_2}{dt} = N_2(a - d(N_1 + N_2) - sb)$$

where a, b, d, and s are positive constants, with $s < 1$ (one can interpret s as a measure of the difference in mortality between the two species).

(b) Find a rescaling of variables that makes the model have just two parameters:

$$\frac{dN_1}{dt} = N_1[\alpha - (N_1 + N_2)]$$

$$\frac{dN_2}{dt} = N_2[\beta - (N_1 + N_2)]$$

(c) Show that the population N_1 and N_2 are related by:

$$N_1(t) = cN_2(t)e^{(s-1)t}$$

for a constant c.

(d) Determine the steady states of the system and their stabilities.

(e) Draw the nullclines of the system, assuming the special values $b = d = 1$, $a = 2$, and $s = 1/2$. Show the directions of movement "South or North" and "East or West" on the N_1 and N_2 nullclines respectively. Then conclude, and show also in the diagram, the directions of movement ("Northeast," etc.) on each region (connected component) of the first quadrant delimited by the nullclines.

Problems ODE3: Chemostat problems

1. For the chemostat model, analyze stability of $\bar{X}_1$ when the parameters are chosen such that the equilibrium $\bar{X}_2$ does not exist.

2. Suppose that we have built a chemostat and we model it as usual, with a Michaelis-Menten growth rate. Moreover, we have measured all constants and, in appropriate units, have:

$$V = 2 \quad \text{and} \quad k_{\max} = \alpha = F = k_n = 1 \,.$$

   What should the concentration of the nutrient in the supply tank be, so that the steady state concentration of bacteria is exactly 20? (Don't worry about units. Assume that all units are compatible.)

   Answering this question should only take a couple of lines. You may use any formula from the notes that you want to.

3. Suppose that we use a Michaelis-Menten growth rate in the chemostat model, and that the parameters are chosen so that a positive steady state exists.

   (a) Show that
   $$N = f(V, F, C_0) = \frac{C_0(F - VK_m) + FK_n}{\alpha\,(F - VK_m)}$$

   and
   $$C = \frac{FK_n}{F - VK_m}$$

   at the positive steady state.

   (b) Show that either of these: (a) increasing the volume of the culture chamber, (b) increasing the concentration of nutrient in the supply tank, or (3) decreasing the flow rate, provides a way to increase the steady-state value of the bacterial population. (Hint: compute partial derivatives.)

4. Suppose that we use $K(C) = kC$ (for some constant k) instead of a Michaelis-Menten growth rate, in the chemostat model.

   (a) Find a change of variables so that only *one* parameter remains.

   (b) Find the steady state(s). Express it (them) in terms of the original parameters. Determine conditions on parameters so that a positive steady state exists, and explain intuitively why these conditions make sense.

   (c) Compare the conditions you just obtained with the ones that we got in the MM case.

   (d) Determine the stability of the positive steady state, if it exists.

5. This is purely a modeling problem (there is no one "correct" answer!). We ask about ways to generalize the chemostat - just provide sets of equations, no need to solve anything.

(a) How would you change the model to allow for two growth-limiting nutrients? Now $K(C_1, C_2)$, the rate of reproduction per unit time, depends on two concentrations.[65] It is a little harder to write how the nutrients are being consumed in this multiple nutrient case. Think about this and be creative.[66]

(b) Suppose that at high densities the microorganism secretes a chemical that inhibits growth. How would you model that?

(c) Model the case when two types of microorganisms compete for the same nutrient.

6. This is yet another variation on the chemostat. Suppose that there is a membrane that filters the outflow, so that the microorganism never flows out (only the nutrient does). Assume, however, that the microorganism dies at a certain rate μN.

(a) Write down a model, assuming $K(C)$ is Michaelis-Menten.

(b) Find a change of variables that leads to a system with three parameters as follows:

$$\frac{dN}{dt} = \alpha_1(\frac{C}{C+1})N - \alpha_3 N$$
$$\frac{dC}{dt} = -(\frac{C}{C+1})N - C + \alpha_2$$

(c) Show that there are two steady states, one with $N = 0$. Show that the second one has positive coordinates provided that:

$$\alpha_1 > \alpha_3 \quad \& \quad \alpha_2 - \frac{\alpha_3}{\alpha_1 - \alpha_3} > 0,$$

(d) Show that this second equilibrium is always stable (if it has positive coordinates).

7. Consider this system of equations (which corresponds to a chemostat with $K(C) = C$):

$$\frac{dN}{dt} = CN - N$$
$$\frac{dC}{dt} = -CN - C + 5.$$

The sketch below has the nullclines (there are vertical arrows on the C axis too).

---

[65]One possibility is to use $K(C_1, C_2) = \max\{K_1(C_1), K_2(C_2)\}$, in the case in which the bacteria decide to use the nutrient that is in most abundance preferentially (this actually happens with certain sugar consumptions - a beautiful example of bacterial computation; search "lac operon" on the web). Another would be to take K as some linear combination of C_1 and C_2, etc. What is the least that one should assume about K, though?

[66]It is amusing to see can be found by typing "*multiple nutrients chemostat*" into Google (you might recognize some of the authors of the first paper that comes up :).
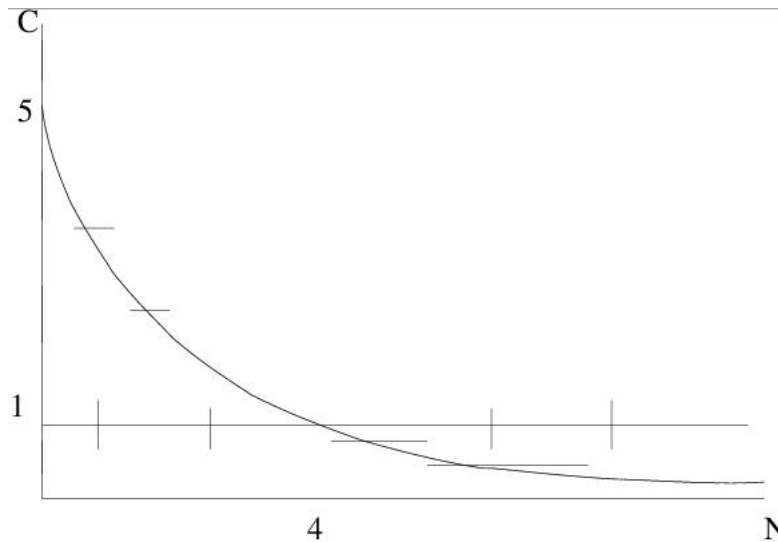
(a) Label the C and N nullclines, and put directions on all arrows. Assign directions to the flows ("NE", "SE", "NW", "SW") in each of the sections of the positive quadrant, partitioned by the nullclines.

(b) Sketch on this diagram a rough plot of the trajectory which starts with a concentration $N = 4$ of bacteria and $C = 2$ of nutrient, and write one or two English sentences explaining what happens to nutrient and bacteria over time (something like: "initially, the nutrient increases and the bacteria decrease, but after a while they both increase, eventually converging to $C = 2$ and $N = 5$").

(c) What is the linearization at the equilibrium $N = 4, C = 1$?

(d) Is the equilibrium $(4, 1)$ stable or unstable? Classify the equilibrium (saddle, spiral, etc). (You are not asked to compute eigenvalues and eigenvectors. It is OK to answer by referring to the trace/determinant plane picture in the notes.)

8. Here is a homework problem involving chemostats with two species and one nutrient. Consider these ODE's:

$$
\begin{aligned}
dB_1/dt &= B_1 f_1(C) - \alpha_1 B_1 \\
dB_2/dt &= B_2 f_2(C) - \alpha_2 B_2 \\
dC/dt &= \beta - \delta C - \frac{1}{\gamma_1} B_1 f_1(C) - \frac{1}{\gamma_2} B_2 f_2(C)
\end{aligned}
$$

for two species of bacteria whose concentrations are B_1 and B_2 and the concentration C of a nutrient. All constants are positive. The functions $f_i(C)$ could be Monod (Michaelis-Menten) or linear.

(a) Interpret the different terms.

(b) [more advanced problem] (i) Suppose $C(0) = \frac{\beta}{\delta}$. Show that then $C(t) \leq \frac{\beta}{\delta}$ for all $t \geq 0$. [Sketch of proof: if this is false, then there are an $\varepsilon > 0$ and a time $t_0 > 0$ such that $C(t_0) = \frac{\beta}{\delta} + \varepsilon$. Without loss of generality, we may assume that $C(t) < \frac{\beta}{\delta} + \varepsilon$ for all $0 \leq t \leq t_0$ (why?). Thus $dC/dt(t_0) < 0$ (why?), but this contradicts that $C(t) < \frac{\beta}{\delta} + \varepsilon$ for all $0 \leq t \leq t_0$ (why?).]

(ii) Suppose now that $f_1\left(\frac{\beta}{\delta}\right) < \alpha_1$. Take a solution with $C(0) = \frac{\beta}{\delta}$ and $B_2(0) = 0$. Prove that $B_1(t) \to 0$ as $t \to \infty$. (If this is too hard, just show that $dB_1/dt(t) < 0$ for all $t \geq 0$.)

This says that the first type of bacteria will become extinct, even if there are no bacteria of the second type. So, from now on we assume that $f_1\left(\frac{\beta}{\delta}\right) > \alpha_1$, and for the same reason, $f_2\left(\frac{\beta}{\delta}\right) > \alpha_2$. Consider these two numbers λ_i, the "break-even concentrations":

$$f_i(\lambda_i) = \alpha_i .$$

Let us pick the smallest of the two; let us say that $\lambda_1 < \lambda_2$ (the other case is entirely analogous). The following theorem is the "competitive exclusion principle" for chemostats:

Theorem: Unless $B_1(0) = 0$, one has that $B_2(t) \to 0$ as $t \to \infty$ and $B_1(t)$ converges to a nonzero steady state, which it is easy to see must be $\frac{\gamma_1}{\alpha_1}(\beta - \delta\lambda_1)$.

This is neat. It says that the bacterium with the smallest requirements completely wins the competition. No co-existence! (The equilibrium with $B_1 = 0$ and $B_2 = \frac{\gamma_2}{\alpha_2}(\beta - \delta\lambda_2)$ is unstable.) There are different theorems that apply to different types of functions f_i; in particular one by Hsu (1978) for Michaelis-Menten and one by Wolkowicz and Lu (1992) for linear f_i's; a nice summary is given in "Competition in the chemostat: Some remarks", by P. De Leenheer, B. Li, and H.L. Smith, Canadian Applied Mathematics Quarterly 11(2003): 229-248 (the theorem is valid more generally for $N > 2$ species as well).

(c) Consider specifically this system:

$$
\begin{aligned}
dB_1/dt &= B_1C - B_1 \\
dB_2/dt &= B_2C - 2B_2 \\
dC/dt &= 3 - C - \frac{1}{2}B_1C - \frac{1}{2}B_2C .
\end{aligned}
$$

(i) Find the equilibrium points of this system. Compute the break-even concentrations λ_1 and λ_2.

(ii) In particular, there will be two equilibria of the forms: $X_1 = (b_1, 0, c_1)$ and $X_2 = (0, b_2, c_2)$. Which of the two should be stable, according to the theory?

(iii) Compute the Jacobians at X_1 and X_2 and find the eigenvalues (use a computer if wanted). Show that all eigenvalues at X_1 are real and negative, but at X_2 there is a real positive eigenvalue.

(d) Suppose that we pick $f_i(C) = C$ as in (c)(iii), but we now have arbitrary parameters α_i, etc. satisfying $f_i\left(\frac{\beta}{\delta}\right) > \alpha_i$.

(i) Compute the equilibria $X_1 = (b_1, 0, c_1)$ and $X_2 = (0, b_2, c_2)$.

(ii) Under what conditions on the parameters does the theory predict that X_2 will be stable?

(iii) Assuming the conditions in (ii), show that the linearized matrix at X_2 is stable and at X_1 is unstable.

9. Consider the steady states of the chemostat, as found in Section 2.2.1. In pharmaceutical and other applications, one wishes to maximize the yield of bacteria at steady state, $\bar{N}_1$. How would you pick values of V, F, C_0 to make this yield large?

10. In Section 2.1.10, we reduced the number of parameters in the chemostat by using $\hat{t} = \frac{V}{F}$, $\hat{C} = k_n$, and $\hat{N} = \dfrac{k_n F}{\alpha V\, k_{\max}}$. Let us now make a different choice for $\hat{C}$ and $\hat{t}$ (leaving $\hat{N}$ the

same), as follows:

$$\hat{t} = \frac{1}{k_{\max}}, \quad \hat{C} = \frac{FC_0}{k_{\max}V} \, .$$

(a) Find the transformed form of the system and group the parameters into two new constants.

(b) Write the stability conditions for the chemostat in terms of the new constants.

(c) Show that $\bar{X}_1$ is stable only when $\bar{X}_2$ is not.

11. For the standard chemostat with Michaelis-Menten kinetics, we found that one condition for the second steady state to be positive was that: $k_{\max} > \frac{F}{V}$.

(a) Prove that, if instead $k_{\max} < \frac{F}{V}$, then $dN/dt < 0$, which means that bacteria will become extinct (no matter how much nutrient there is!).

(b) Interpret the second condition, $C_0 > \frac{k_n}{\frac{V}{F}k_{\max}-1}$.

Problems ODE4: Chemotherapy, metabolism, and drug infusion problems

1. Consider the chemotherapy model (section "Effect of Drug on Cells in an Organ"). Suppose that $K(C)$ is Michaelis-Menten.

   (a) Show how to reduce the model to having just three constants.

   (b) There are again two steady states, one with $N = 0$. Find conditions under which there is a second one that has positive coordinates. Interpret biologically what your conditions mean.

   (c) In contrast to the chemostat (where the objective is to get the microorganisms to grow), it would be desirable if the equilibrium with $N = 0$ is stable and the second one either doesn't exist (in the positive quadrant) or is unstable. Why would a stable second equilibrium be bad? (Just one sentence, please.)

   (d) Find conditions guaranteeing that the equilibrium with $N = 0$ is stable and show that, under these conditions, the second equilibrium, if it is in the first quadrant, must be unstable.

2. For the chemotherapy model discussed above, write a computer program in your favorite package (MATLAB, Mathematica, Maple, whatever) to simulate it (or use "Jode"), and plot solutions from several different initial conditions (show $N(t)$ and $C(t)$ versus t on the same plot, using different line styles such as lines and dots or dashes or different colors). Do so for sets of parameters that illustrate a few of the possible cases (second equilibrium exists as a positive solution or now; equilibrium with $N = 0$ is stable or not)

3. For the chemotherapy model discussed above, and assuming that parameters are so that the equilibrium with $N = 0$ is stable and the one in the positive orthant is unstable, there are two possibilities: (a) all solutions (except for a set of measure zero of initial conditions) converge to the equilibrium with $N = 0$ is stable; (b) there is set of initial conditions with nonzero measure for which solutions do not converge to the equilibrium with $N = 0$. Determine which is the case, and prove your conclusions rigorously. (Some numerical experimentation will be useful, of course.)

4. We base this problem on the following paper:

   M.B. Rabinowitz, G.W. Wetherill, and J.D. Kopple, "Lead metabolism in the normal human: stable isotope studies," in *Science*, vol. 182, 1973, pp. 725 - 727.

   as well as a writeup from the Duke Connected Curriculum Project (by L.C. Moore and D.A. Smith) (we use the wording from this writeup):

   Lead enters the human body from the environment by inhalation, by eating, and by drinking. From the lungs and gut, lead is taken up by the blood and rapidly distributed to the liver and kidneys. It is slowly absorbed by other soft tissues and very slowly by the bones. Lead is excreted from the body primarily through the urinary system and through hair, nails, and sweat:
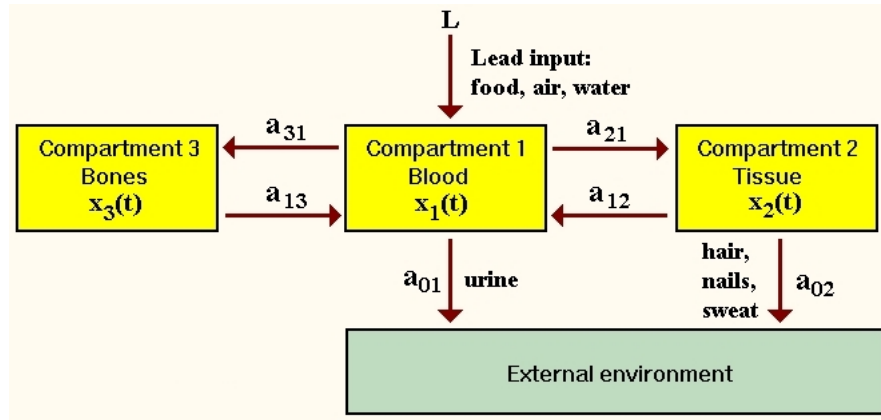
We model the flow of lead into, through, and out of a body with separate compartments for blood, bones, and other tissues, plus a compartment for the external environment. For $i = 1, 2, 3$, we let $x_i(t)$ be the amount of lead in compartment i at time t, and we assume that the rate of transfer from compartment i to compartment j is proportional to $x_i(t)$ with proportionality constant a_{ji}.

We assume that exposure to lead in the environment results in ingestion of lead at a constant rate L.

The units for amounts of lead are micrograms (μg), and time t is measured in days.

Rabinowitz et. al. (paper cited above) measured over an extended period of time the lead levels in bones, blood, and tissue of a healthy male volunteer living in Los Angeles. Their measurements produced the following transfer coefficients for movement of lead between various parts of the body and for excretion from the body. Note that, relatively speaking, lead is somewhat slow to enter the bones and very slow to leave them. The estimated rates are (units are days$^{-}1$):

$$a_{21} = 0.011, a_{12} = 0.012$$

(from blood to tissue and back),

$$a_{31} = 0.0039, a_{13} = 0.000035$$

(from blood to bone and back), and

$$a_{01} = 0.021, a_{02} = 0.016$$

(excretion from blood and tissue), and they estimated that the average rate of ingestion of lead in Los Angeles over the period studied was $L = 49.3 \mu g$ per day.

(a) Find the steady state of the corresponding system. (You need to solve a set of three linear equations.)

(b) Now repeat the problem assuming that the coefficient for blood to tissue transfer is ten times bigger: $a_{21} = 0.11$.

(c) Find the steady state of the modified system, and conclude that in steady state the amount of lead in the tissue is about three times higher than in the original model.

(d) Use a computer (MATLAB, Maple, Mathematica, JOde, whatever you prefer) to plot the amount of lead in the tissue, starting from zero initial conditions, for the original ($a_{21} = 0.011$) model, for two years. Notice that even after two years, the amount is not quite near steady state (it is about 620 μg). What is it after 100 years?

Problems ODE5: Epidemiology problems

1. For the SIRS model, suppose that $\beta = \nu = \gamma = 1$. For what values of N does one have stable spirals and for what values does one get stable nodes, for $\bar{X}_2$?

2. Take the SIRS model, and suppose that the parameters are so that a positive steady state exists. Now assume that a new medication is discovered, which multiplies by 20 the rate at which people get cured (that is, become "removed" from the infectives). However, at the same time, a mutation in the virus which causes this disease makes the disease 5 times as easily transmitted as earlier. How does the steady state number of susceptibles change?

   (The answer should be stated as something like "it is doubled" or "it is cut in half".)

   Answering this question should only take a couple of lines. You may use any formula from the notes that you want to.

3. We consider an SIR model with vital dynamics. A per-capita mortality rate δ is assumed to be the same for all types of individuals, so that there will be terms "$-\delta S$" and so forth for each of the three classes. We also assume that births lead to new susceptibles at a rate δ which is identical to the mortality rate, so we model the "birth" rate as a term "$\delta(S+I+R)$" added to dS/dt. (Of course, all these assumptions are a bit artificial. We make them in order to have a simpler model.) The figure shown below illustrates the basic rates (using δN to indicate $\delta(S+I+R)$).
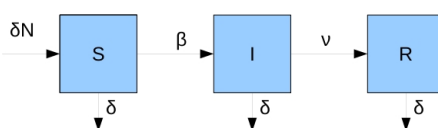
   (a) Write a set of equations describing the model.

   (b) Find the steady states of the system and their stabilities.

   (c) Under what conditions does the epidemic occur, i.e., what conditions on the parameters allow a steady state with positive coordinates?

4. We continue with the previous problem, but we now consider an SIRS model: there is a flow from R back to S, as shown in the figure.
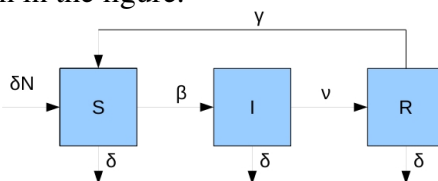
   (a) Write a set of equations describing the model.

   (b) Find the steady states of the system and their stabilities.

   (c) Under what conditions does the epidemic occur, i.e., what conditions on the parameters allow a steady state with positive coordinates?

5. In the following model, we allow the emigration of susceptibles:

$$
\begin{aligned}
\frac{dS}{dt} &= -g(I)S - \lambda S \\
\frac{dI}{dt} &= g(I)S - \gamma I \\
\frac{dR}{dt} &= \lambda S + \gamma I
\end{aligned}
$$

with $g(x) = xe^{-x}$.

(a) Interpret the various terms in the equations.

(b) Show that the epidemic will always tend to extinction, in the sense that both infectives and susceptibles converge to zero.

6. The following model describes a directly transmitted viral microparasite:

$$
\begin{aligned}
\frac{dS}{dt} &= bN - \beta SI - bS \\
\frac{dI}{dt} &= \beta SI - rI - bI \\
\frac{dR}{dt} &= rI - bR.
\end{aligned}
$$

The variables and S, I and R represent the numbers of susceptible, infective and immune individuals, and b, β, r and N are all positive constants. (We may think of b as a death rate, caused by something other than the parasite, so that all rates are the same.)

Show that, the initial state $(N, 0, 0)$ is a steady state.

Find the Jacobian matrix (a 3×3 matrix) for the linearization at $(N, 0, 0)$ and compute its eigenvalues.

Now find a number (which we think of as a "threshold population size") N_c with the property that, if $N < N_c$, then $(N, 0, 0)$ is stable.

(The interpretation is that, for such N, the parasite cannot maintain itself in the population, and both the infective and the immune class eventually die out.)

7. Let us consider a "criss-cross" venereal infection model, in which the removed class is permanently immune. We assume the following influences, employing the usual notations for the susceptible, infective, and removed classes:
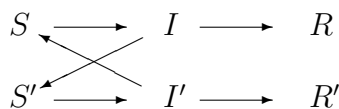
(I' infects S, and I infects S').

(a) Explain why this is a good model:

$$\frac{dS}{dt} = -rSI'$$

$$\frac{dS'}{dt} = -r'S'I$$

$$\frac{dI}{dt} = rSI' - aI$$

$$\frac{dI'}{dt} = r'S'I - a'I'$$

$$\frac{dR}{dt} = aI$$

$$\frac{dR'}{dt} = a'I'$$

(the parameters are all positive).

Let the initial values for S, I, R, S', I' and R' be S_0, I_0, 0 and S_0', I_0', 0 respectively.

(b) Show that the female and male populations stay constant, and therefore $S(t) = S_0 \exp[-rR'/a']$. Conclude that $\lim_{t\to\infty} S(t) > 0$ and $\lim_{t\to\infty} I(t) = 0$. Deduce similar results for S' and I'.

(c) Obtain an equation which determines $S(\infty)$ and $S'(\infty)$.

(d) Show that a condition for an epidemic to occur is at least one of:

$$\frac{S_0 I_0'}{I_0} > a/r, \quad \frac{S_0' I_0}{I_0'} > a'/r'.$$

Hint: Think of $\frac{dI}{dt}$ and $\frac{dI'}{dt}$ at $t = 0$.

(e) What single condition would ensure an epidemic?

8. As in the notes, we study a virus that can only be passed on by heterosexual sex. There are two separate populations, male and female: we use $\bar{S}$ to indicate the susceptible males and S for the females, and similarly for I and R.

The equations analogous to the SIRS model are:

$$\frac{d\bar{S}}{dt} = -\bar{\beta}\bar{S}I \qquad + \bar{\gamma}\bar{R}$$

$$\frac{d\bar{I}}{dt} = \bar{\beta}\bar{S}I - \bar{\nu}\bar{I}$$

$$\frac{d\bar{R}}{dt} = \bar{\nu}\bar{I} - \bar{\gamma}\bar{R}$$

$$\frac{dS}{dt} = -\beta S\bar{I} \qquad + \gamma R$$

$$\frac{dI}{dt} = \beta S\bar{I} - \nu I$$

$$\frac{dR}{dt} = \nu I - \gamma R.$$

This model is a little difficult to study, but in many STD's (especially asymptomatic), there is no "removed" class, but instead the infecteds get back into the susceptible population. This gives:

$$
\begin{aligned}
\frac{d\bar{S}}{dt} &= -\bar{\beta}\bar{S}I && + \bar{\nu}\bar{I} \\
\frac{d\bar{I}}{dt} &= \bar{\beta}\bar{S}I - \bar{\nu}\bar{I} \\
\frac{dS}{dt} &= -\beta S\bar{I} && + \nu I \\
\frac{dI}{dt} &= \beta S\bar{I} - \nu I \,.
\end{aligned}
$$

Writing $\bar{N} = \bar{S}(t) + \bar{I}(t)$ and $N = S(t) + I(t)$ for the total numbers of males and females, and using these two conservation laws, we then concluded that one may just study the following set of two ODE's:

$$
\begin{aligned}
\frac{d\bar{I}}{dt} &= \bar{\beta}(\bar{N} - \bar{I})I - \bar{\nu}\bar{I} \\
\frac{dI}{dt} &= \beta(N - I)\bar{I} - \nu I \,.
\end{aligned}
$$

Parts (a)-(c) refer to this reduced model.

(a) Prove that there are two equilibria, the first of which is $I = \bar{I} = 0$ and a second one, which exists provided that:

$$
R_0\bar{R}_0 = \left(\frac{N\beta}{\nu}\right)\left(\frac{\bar{N}\bar{\beta}}{\bar{\nu}}\right) > 1
$$

and is given by $I = \frac{N\bar{N} - (\nu\bar{\nu})/(\beta\bar{\beta})}{\nu/\beta + \bar{N}}$, $\bar{I} = \frac{N\bar{N} - (\nu\bar{\nu})/(\beta\bar{\beta})}{\bar{\nu}/\bar{\beta} + N}$.

(b) Prove that the first equilibrium is unstable, and the second one stable.

(c) What vaccination strategies could be used to eradicate the disease?

(d) Now consider the full model (six dimensional, with removeds). How many linearly independent conservation laws are there?

(e) Again for the full model. Reduce by conservation to a system of 5 or less equations (how many, depends on how many conservation laws you found in (d)). Pick some set of numerical parameters (any you want) such that $R_0\bar{R}_0 = 2$. Determine, using computer simulations, what the solutions look like. (You may be able to find the steady states algebraically, too.)

For your answer, attach some plots of solutions $I(t)$ and $\bar{I}(t)$ as a function of time.

Problems ODE6: Chemical kinetics problems

1. Suppose that the enzyme E can react with the substrate A in such a way that up to two copies of A may bind to E at the same time. We model this by the following chemical network:

$$A + E \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} C_1 \overset{k_2}{\longrightarrow} P + E$$

$$A + C_1 \underset{k_{-3}}{\overset{k_3}{\rightleftarrows}} C_2 \overset{k_4}{\longrightarrow} P + C_1$$
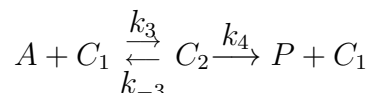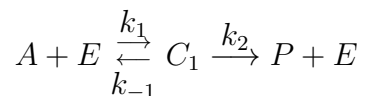
where the k's are the rate constants, P a product of the reaction, and C_1 and C_2 are enzyme-substrate complexes.

(a) What are the species vector S and the reaction vector $R(S)$?

(b) Find the stoichiometry matrix Γ and calculate its rank.

(c) Compute the product $\Gamma R(S)$, and write down the differential equation model based on mass action kinetics.

(d) Find a conservation law, assuming that $a(0) = a_0$, $e(0) = e_0$, $c_1(0) = c_2(0) = p(0) = 0$.

(d') Find a basis of the left nullspace of Γ and compare it with your answer of part (d).

(e) Use part (d) and write a system of equations for just a, c_1 and c_2.

(f) If $\varepsilon = \frac{e_0}{a_0} \ll 1$, $\tau = k_1 e_0 t$, $u = \frac{a}{a_0}$, $\nu_i = \frac{c_i}{e_0}$ for $i = 1, 2$, show that the reaction mechanism reduces to the following form in terms of the new variables (provide expressions for f and g):

$$\begin{aligned} \frac{du}{d\tau} &= f(u, \nu_1, \nu_2) \\ \varepsilon \frac{d\nu_i}{d\tau} &= g_i(u, \nu_1, \nu_2), \quad i = 1, 2. \end{aligned}$$

2. Consider the following chemical reaction network, which involves 4 substances called M, E, C, P:

$$3M + 2E \underset{1}{\overset{1}{\rightleftarrows}} C \overset{1}{\longrightarrow} 2E + P.$$
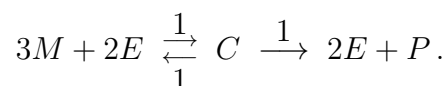
(a) Find the species vector S and the reaction vector $R(S)$ (assuming mass action kinetics).

(b) Find the stoichiometry matrix Γ.

(c) Compute the product $\Gamma R(S)$ and show the set of 4 differential equations for M, E, C, P.

(d) Find the rank of Γ.

(e) What is the dimension of the left nullspace of Γ?

(f) Find a basis of the left nullspace of Γ (conservation laws).

3. We discussed the chemical kinetics formulation of an example that may be represented as in Figure 2.1(a).

Many cell signaling processes involve double instead of single transformations such as addition of phosphate groups. A model for a double-phosphorylation as in Figure 2.1(b) corresponds to

Figure 2.1: *(a) One-step*   *and (b) two-step transformations*

reactions as follows (we use double arrows for simplicity, to indicate reversible reactions):

$$E + S_0 \;\leftrightarrow\; ES_0 \;\to\; E + S_1 \;\leftrightarrow\; ES_1 \;\to\; E + S_2$$
$$F + S_2 \;\leftrightarrow\; FS_2 \;\to\; F + S_1 \;\leftrightarrow\; FS_1 \;\to\; F + S_0 \tag{2.19}$$
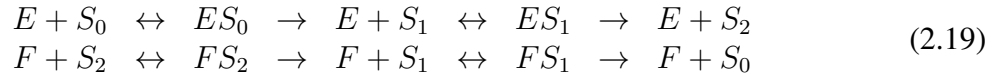
where "ES_0" represents the complex consisting of E bound to S_0 and so forth. For simplicity, assume that all constants are equal to 1.

(a) Find the stoichiometry matrix and write a corresponding system of ODE's.

(b) Show that there is a basis of conservation laws consisting of three vectors.

4. In the quasi-steady state derivations, suppose that, instead of $e_0 \ll s_0$, we know only the weaker condition:
$$e_0 \ll (s_0 + K_{\mathrm{m}}) \,.$$

Show that the same formula for product formation is obtained. Specifically, now pick:
$$x = \frac{s}{s_0 + K_{\mathrm{m}}}\,, \quad y = \frac{c}{e_0}\,, \quad \varepsilon = \frac{e_0}{s_0 + K_{\mathrm{m}}}$$

and show that the equations become:

$$\frac{dx}{dt} \;=\; \varepsilon \Big[k_{-1}\,y \;-\; k_1\,(s_0 + K_{\mathrm{m}})\,x\,(1 - y) \Big]$$
$$\frac{dy}{dt} \;=\; k_1 \Big[(s_0 + K_{\mathrm{m}})\,x \;-\; \big(K_{\mathrm{m}} + (s_0 + K_{\mathrm{m}})\,x\big)\,y \Big] \,.$$

Now set $\varepsilon = 0$. In conclusion, one doesn't need $e_0 \ll s_0$ for the QSS approximation to hold. It is enough that K_{m} be very large, that is to say, for the rate of formation of complex k_1 to be very small compared to $k_{-1} + k_2$ (sum of dissociation rates).

5. As in the text, we consider a simplification of allosteric inhibition in which binding of substrate can always occur, but product can only be formed (and released) if I is not bound. In addition, we will also assume that binding of S or I to E are independent of each other. (If we don't assume this, the equations are still the same, but we need to introduce some more kinetic constants k's.)

A reasonable chemical model is, then:

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

$$EI + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} EIS$$

$$E + I \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} EI$$

$$ES + I \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} EIS$$
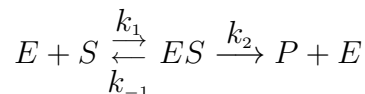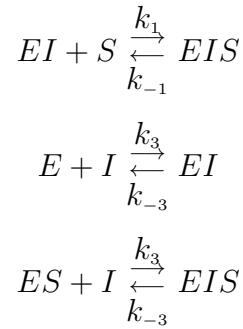
where "EI" denotes the complex of enzyme and inhibitor, etc.

Prove that there results under quasi-steady state approximation a rate

$$\frac{dp}{dt} = \frac{V_{\text{max}}}{1 + i/K_{\mathbf{i}}} \cdot \frac{s^2 + as + b}{s^2 + cx + d}$$

for some suitable numbers $a = a(i), \dots$ and a suitably defined $K_{\mathbf{i}}$.

6. A process in which a chemical is involved in its own production is called *autocatalysis*. A simple example of autocatalysis is:

$$E + X \underset{k_1}{\overset{k_{-1}}{\rightleftharpoons}} 2X$$
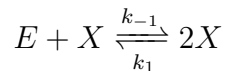
in which a molecule of X combines with a molecule of E to produce two molecules of X. To simplify, we will assume that E has a constant concentration, which we denote as e. This might make sense if E is very large, so it is not consumed in any appreciable manner during the reactions, or perhaps if it is beingh replentished externally during the process. In that case, we really can look at the simpler model

$$X \underset{k_{-1}}{\overset{k_1 e}{\rightleftharpoons}} 2X$$
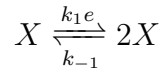
Denote the concentration of X by x. Now answer the following questions:

(a) Write a scalar ODE for $x(t)$ that describes the system (assuming mass action kinetics). (Suggestion: first find Γ and $R(S)$, which are a 2×1 and 1×2 matrix respectively.)

(b) Find the steady states and their stabilities.

(c) Show that $x(t) \to \frac{k_1 e}{k_{-1}}$ as $t \to \infty$.

7. We consider a simple model of gene expression, in which there is also another species, a transcription factor (which we will denote by R) that can bind to DNA at the promoter site of a gene, and produce a complex (which we will denote by C). We assume that transcription is also possible from this "occupied promoter" C:

$$D \xrightarrow{\alpha} D + M, \quad M \xrightarrow{\beta} 0, \quad M \xrightarrow{\theta} M + P, \quad P \xrightarrow{\delta} 0, \quad D + R \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} C, \quad C \xrightarrow{\varepsilon} C + M.$$
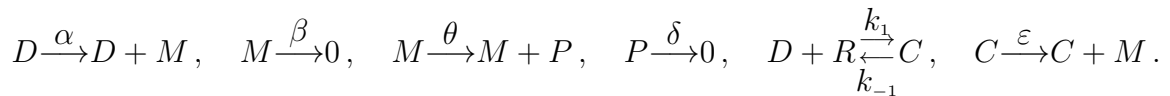
(a) Write the stoichiometry matrix, reaction vector, and differential equations.

(b) Answer: if $\varepsilon < \alpha$, is R an enhancer (activator) or a repressor?

(c) Answer: if $\varepsilon > \alpha$, is R an enhancer (activator) or a repressor?

(d) Assume now that R is an enhancer. Suppose that you can do some genetic engineering and change k_{-1}. Answer: how would you change this parameter (should k_{-1} be incresed or decreased?) so that the effect of R is magnified?

(e) Now consider the following special parameters: $\alpha = \beta = \theta = \delta = k_{-1} = 1$ and $k_1 = 3$. Find the steady state that satisfies that $D + C = 2$ and $R + C = 2$. (It is easy to see that there are infinitely many steady states, but a unique one once that we impose constraints on $D+C$ and $R+C$.) Your answers for the steady state values will involve some expressions that have ε in them, such as "$\frac{2\varepsilon+1}{5}$".

8. Consider the following system:

$$A + X \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} 2X$$

$$B + X \overset{k_2}{\longrightarrow} C$$
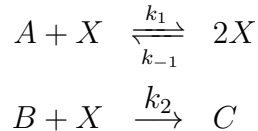
where a molecule of X combines with a molecule of A to produce two molecules of X and a molecule of X is combines with a molecule of B to produce molecule of C. As in a previous problem, supposing A and B have constant concentrations a and b, we will look at the simpler system:

$$X \underset{k_{-1}}{\overset{k_1 a}{\rightleftharpoons}} 2X$$

$$X \overset{k_2 b}{\longrightarrow} C$$
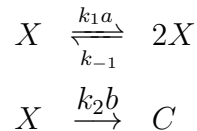
We denote the concentration of X as x. Write explicitly the vector S, stoichiometry matrix Γ, and $R(S)$, as well as the ODEs for X and for C.

9. In the competitive inhibition model we ended up with these equations for dc_1/dt and dc_2/dt:

$$\frac{dc_1}{dt} = k_1 s(e_0 - c_1 - c_2) - (k_{-1} + k_2)c_1$$

$$\frac{dc_2}{dt} = k_3 i(e_0 - c_1 - c_2) - k_{-3}c_2 .$$

(a) The notes claim that formally setting $dc_1/dt = 0$ and $dc_2/dt = 0$ gives:

$$c_1 = \frac{K_i e_0 s}{K_m i + K_i s + K_m K_i}, \quad c_2 = \frac{K_m e_0 i}{K_m i + K_i s + K_m K_i}$$

where $K_m = \frac{k_{-1}+k_2}{k_1}$ and $K_i = \frac{k_{-3}}{k_3}$. Prove these formulas for c_1, c_2 (that is, solve for c_1, c_2 the two equations obtained by setting the right-hand sides of dc_1/dt and dc_2/dt to zero).

(b) It is also stated that, with $V_{\max} = k_2 e_0$, we obtain

$$\frac{dp}{dt} = \frac{V_{\max} s}{s + K_m(1 + i/K_i)} .$$

Prove this.

(c) Now take the values $V_{\max} = 1$, $K_i = K_m = 1$, and $i = 10$. Plot dp/dt as a function of s for $s \in [0, 1000]$.

(d) Repeat the above for larger inhibitor, $i = 100$. Notice that, as discussed in class, the graph has changed little.

(e) Now take $V_{\max} = 1$, $K_{\mathrm{m}} = 1$, and $i = 100$, but make much smaller: $K_{\mathrm{i}} = 0.001$. Plot and observe that the graph stays at about 1/1000 of its previous value.

Think about what it means to say that K_{i} is small.

10. This problem refers to the simplest model of gene expression treated in the notes, for which the network of reactions is:

$$D \xrightarrow{\alpha} D + M, \quad M \xrightarrow{\beta} 0, \quad M \xrightarrow{\theta} M + P, \quad P \xrightarrow{\delta} 0.$$
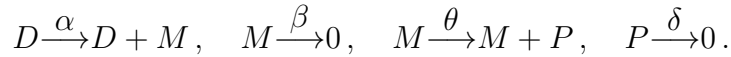
The set of ODE's is:

$$\frac{dM}{dt} = \alpha D - \beta M$$
$$\frac{dP}{dt} = \theta M - \delta P$$

together with $dD/dt = 0$. Since $D(t)$ is constant, let us write $D(t) = d_0$ for all t. Redefining α as αd_0, we will assume from now on that $D = 1$.

(a) Find the steady state of this system (there is only one).

(b) Show (by plugging-into the equations) that this is a solution, for any two constants c_1 and c_2:

$$M(t) = \left(e^{-\beta t} c_2 - \frac{\alpha \theta}{\beta (\beta - \delta)} \right) \frac{\delta - \beta}{\theta}$$
$$P(t) = e^{-\beta t} c_2 - \frac{\alpha \theta}{\beta (\beta - \delta)} + e^{-\delta t} c_1 + \frac{\alpha \theta}{\delta (\beta - \delta)}$$

(c) Compute the limit of $P(t)$ as $t \to \infty$ using the expression given above.

Problems ODE7: Multiple steady states, sigmoidal responses, ultrasensitivity

1. Do this problem:

   http://www.math.rutgers.edu/~sontag/JODE/gardner_cantor_collins_toggle.html

2. This problem deals with the material on cell differentiation. We consider a toy "1-d organism", with cells are arranged on a line Each cell expresses a certain gene X according to the same differential equation

$$\frac{dx}{dt} = f(x) + a$$

   but the cells toward the left end receive a low signal $a \approx 0$, while those toward the right end see a high signal a (and the signal changes continuously in between). The level of expression starts at $x(0) = 0$ for every cell.

   This is what $f + a$ looks like, for low, intermediate, and high values of a respectively:
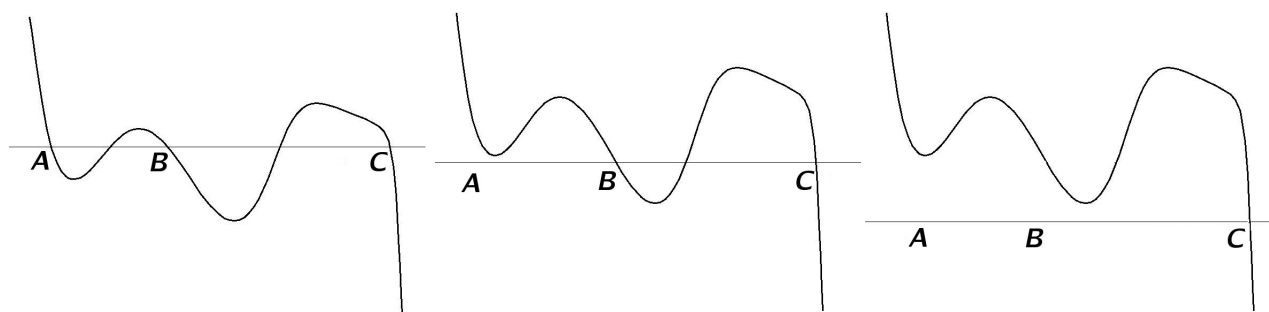
   We let the system settle to steady state.

   After the system has so settled, we next suddenly change the level of the signal a, so that from now on *every* cell sees the *same* value of a. The value of a that every cell is exposed to, in the second part of the experiment, corresponds to an intermediate value that gives a graph like the second (right) one above.

   Like in the example worked out above, we ask what the patterns will be after the first and second experiments.

   Here are a few possibilities of what will be seen after the first and the second parts of the experiment. Circle the correct one (no need to explain).

   (a) 000000000000 → AAAABBBBCCCC → AAAAAABBBBBB

   (b) 000000000000 → AAAAAABBBBBB → BBBBBBBBBBBB

   (c) 000000000000 → AAAAAAAAAAAA → BBBBBBBBBBBB

   (d) 000000000000 → BBBBAAAACCCC → AAAAAACCCCCC

   (e) 000000000000 → AAAABBBBCCCC → BBBBCCCCCCCC

   (f) 000000000000 → AAAAAABBBBBB → AAAAAABBBBBB

   (g) 000000000000 → AAAABBBBCCCC → CCCCCCAAAAAA

   (h) 000000000000 → AAAABBBBCCCC → AAAAAABBBBBB

   (i) 000000000000 → AAAABBBBCCCC → BBBBBBBBCCCC

   (j) 000000000000 → AAAABBBBCCCC → CCCCCCCCCCCC

(k) 000000000000 → CCCCCCCCCCCC → BBBBBBBBBBBB

These next few problems deal with steady-states as function of input: hyperbolic and sigmoidal responses, adaptation:

3. Consider this reaction:

$$A \underset{1}{\overset{s}{\rightleftarrows}} B$$
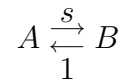
which describes for example a phosphorylation of A into B with rate constant s, and a reverse de-phosphorylation with rate constant 1.

One may think of s as the concentration of an enzyme that drives the reaction forward.

   (a) Write equations for this reaction (assuming mass action kinetics; for example $da/dt = -sa + b$).

   (b) Observe that $a(t) + b(t)$ is constant. From now on, assume that $a(0)=1$ and $b(0)=0$. What is the constant, then?

   (c) (Still assuming $a(0)=1$ and $b(0)=0$.) Use the conservation law from (b) to eliminate a and write just one equation for b.

   (d) Find the steady state $b(\infty)$ of this equation for b and think of it as a function of s. Answer this: is $b(\infty)$ a hyperbolic or sigmoidal function of s?

   (e) Find the solution $b(t)$ (this is just to practice ODE's).

4. Consider this reaction:

$$A \underset{1}{\overset{s}{\rightleftarrows}} B \underset{1}{\overset{s}{\rightleftarrows}} C$$
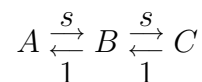
which describes for example a phosphorylation of A into B, and then of B into C, with rate constant s, and reverse de-phosphorylations with rate constant 1.

   (a) Write equations for this reaction (assuming mass action kinetics).

   (b) Find a conservation law, assuming that $a(0)=1$ and $b(0)=c(0)=0$, and, using this law, eliminate b and write a system of equations for just a and c.

   (c) Find the steady state $(a(\infty), c(\infty))$ of this equation for a, c and think of $c(\infty)$ as a function of s. Answer this: is $c(\infty)$ a hyperbolic or sigmoidal function of s?

5. Consider this reaction:
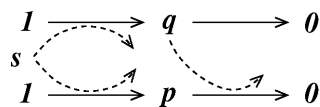
where the dashed lines mean that s and q do not get consumed in the corresponding reactions (they both behave as enzymes). There are also constants k_i for each of the rates (not shown).

Make sure that you understand then why these are the reasonable mass-action equations to describe the system:

$$\frac{dp}{dt} = k_1 s - k_2 pq$$

$$\frac{dq}{dt} = k_3 s - k_4 q$$

(or one could have used, instead, a more complicated Michaelis-Menten model).

(a) Find the steady state, written as a function of s.

(b) Note that $p(\infty)$ (though not $q(\infty)$) is independent of s.

This is an example of *adaptation*, meaning that the system transiently responds to a "signal" s (assumed a constant), but, after a while, it returns to some "default" value which is independent of the stimulus s (and hence the system is ready to react to other signals).

(c) Graph (using for instance JOde) the plot of $p(t)$ versus t, assuming that $k_1=k_2=2$ and $k_3=k_4=1$, and $p(0)=q(0)=0$, for each of the following three values of s: $s = 0.5, 3, 20$.

You should see that $p(t) \to 1$ as $t \to \infty$ (which should be consistent to your answer to part (b)) but that the system initially reacts quite differently (in terms of "overshoot") for different values of s.

6. This problem refers to the Goldbeter-Koshland model. Let $G(r)$ be the solution x of equation (2.18): $r\dfrac{x}{K + x} = \dfrac{1 - x}{L + 1 - x}$ as a function of r (recall that r is the ratio between enzymes and x is the fraction of unmodified substrate, or equivalently, $1 - x$ is the modified fraction).

(a) Let us assume that $K = L = \varepsilon$ and that $r \neq 1$. Show that this choice of G:

$$G(r) = \frac{(1 - r - \varepsilon - \varepsilon r) + \sqrt{(1 - r - \varepsilon - \varepsilon r)^2 + 4(1 - r)\varepsilon}}{2(1 - r)}$$

results in a function G which has the property that $G(r) > 0$ for all r. You will need to separately analyze the case when the denominator is positive or negative, depending on whether $r < 1$ or $r > 1$. The rest of the problem uses this formula for G. (One can also prove that $G(r) < 1$ when picking the positive root, and that the negative root would not have the properties that both $G(r) > 0$ and $G(r) < 1$ for all r.)

(b) Show that $\lim\limits_{\varepsilon \to 0} G(r) = \dfrac{1}{2} + \dfrac{1}{2}\dfrac{|1 - r|}{1 - r}$.

(c) Conclude from the above that $\lim_{\varepsilon \to 0} G(r) = 1$ if $r < 1$ and $\lim_{\varepsilon \to 0} G(r) = 0$ if $r > 1$.

(d) Use the above to sketch (not using a computer!) the graph of $G(r)$ for $r > 0$, assuming that ε is very small.

(e) We said that for K, L not too small, we should have a hyperbolic-looking as opposed to a sigmoidal-looking plot. Use a computer or graphing calculator to plot $G(r)$ when $K = L = 1$. Include a printout of your plot with your answers.

Problems ODE8: Oscillations and excitable systems

1. For the van der Pol oscillator, show that:

    (a) There are no periodic orbits contained entirely inside the half-plane $\{(x, y), x > 1\}$.

    (b) There are no periodic orbits contained entirely inside the half-plane $\{(x, y), x < -1\}$.

   (Use Bendixon's criterion to rule out such orbits.)

2. Consider a system with equations as follows, where we assume that $0 < \varepsilon \ll 1$:

$$
\begin{aligned}
dx/dt &= f(x) - y \\
dy/dt &= \varepsilon(g(x) - y).
\end{aligned}
$$

   Consider these four possibilities for the nullclines $y = f(x)$ and $y = g(x)$:
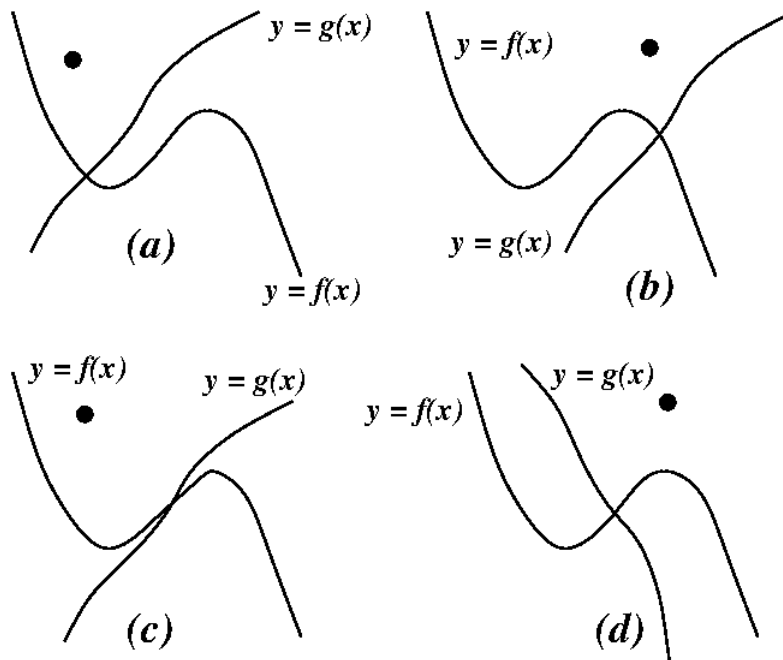
    (i) What can you say about stability of the steady states in each case?

    (ii) Sketch directions of movement in each figure, making sure to show where the vector field is "almost horizontal". Include arrows on all nullclines, too.

    (iii) For each of these, sketch trajectories when starting from the points labeled by the dark circle.

    (iv) Sketch $x(t)$ and $y(t)$ as a function of t, for each of the examples, when starting from the points labeled by the dark circle.

3. Consider the system

$$
\begin{aligned}
dx_1/dt &= \mu x_1 - \omega x_2 - x_1(x_1^2 + x_2^2) \\
dx_2/dt &= \omega x_1 + \mu x_2 - x_2(x_1^2 + x_2^2)
\end{aligned}
$$

which we analyzed in polar coordinates. Now represent the pair (x_1, x_2) by the complex number $z = x_1 + ix_2$. Show that the equation becomes, for $z = z(t)$:

$$dz/dt = (\mu + \omega i)z - |z|^2 z.$$

(Hint: use that $dz/dt = dx_1/dt + idx_2/dt$.)

4. Check for which of the following vector fields

$$F(x, y) = \left(\begin{array}{c} f(x, y) \\ g(x, y) \end{array} \right)$$

is the unit square $[0, 1] \times [0, 1]$ a trapping region (so that one may attempt to apply Poincaré-Bendixson).

(a) $f(x, y) = y - x, \quad g(x, y) = x - y$

(b) $f(x, y) = y(x - 1), \quad g(x, y) = -x$

(c) $f(x, y) = x(y - 1), \quad g(x, y) = -y$

(d) $f(x, y) = \cos \pi(x + y/8), \quad g(x, y) = x^2 - y^2$

(e) $f(x, y) = xy(x - 1), \quad g(x, y) = -xy - y$

(f) $f(x, y) = xy(x - 1), \quad g(x, y) = -xy - 1$

You must check, for each case, in which directions the vector fields point at each of the four sides of the unit square.

5. We consider this system of differential equations:

$$\frac{dx}{dt} = f(x, y)$$
$$\frac{dy}{dt} = g(x, y)$$

and want to know if the region formed by intersecting the unit disk $x^2 + y^2 \leq 1$ with the upper half-plane is a trapping region.
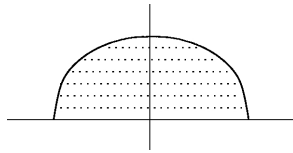
Answer yes or no for each (*justify your answer by computing the dot products with an appropriate normal vector*) for each of these cases:

(a) $f(x, y) = -x, \quad g(x, y) = -y^2$

(b) $f(x, y) = y, \quad g(x, y) = -x$

(c) $f(x, y) = -xy^2 - x, \quad g(x, y) = x^2 y$

6. We had shown that the van der Pol oscillator has a periodic orbit somewhere inside the hexagonal region bounded by the segments $x = 3, -3 \leq y \leq 6$, etc. We now want to improve our estimate and claim that there is a periodic orbit that is not "too close" to the origin. Show that there is a periodic orbit in the region which lies between the circle of radius $\sqrt{3}$ and the hexagonal region that we considered earlier. (You need to consider which way the vector field points when one is on the circle.)

7. In each of the figures below, we show nullclines as well as a steady state (indicated with a large solid dot) which is assumed to be repelling. You should draw, in each figure, a trapping region that contains this point. (Which allows us to conclude that there is a periodic orbit.)
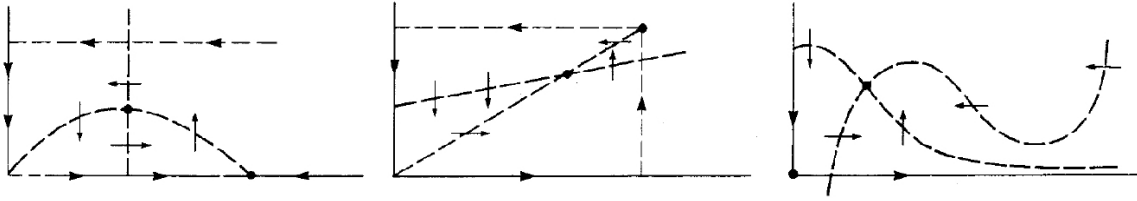
8. Glycolysis is a metabolic pathway that converts glucose into pyruvate, in the process making ATP as well as NADH. One of the intermediate reactions involves ADP and F6P (Fructose-6-phosphate), of which a simple model is given by:

$$dx/dt = -x + ay + x^2 y$$
$$dy/dt = b - ay - x^2 y$$

where x, y denote are concentrations of ADP and F6P, and a, b are two positive constants.

(a) Plotted below are the nullclines as well as a possible trapping region. Show that the arrows on the boundary of the trapping region are really as shown, i.e. that they point to the inside. (The only nontrivial part is dealing with the diagonal, upper-right, boundary. It has slope -1. You will use that $x \geq b$ on this line.)
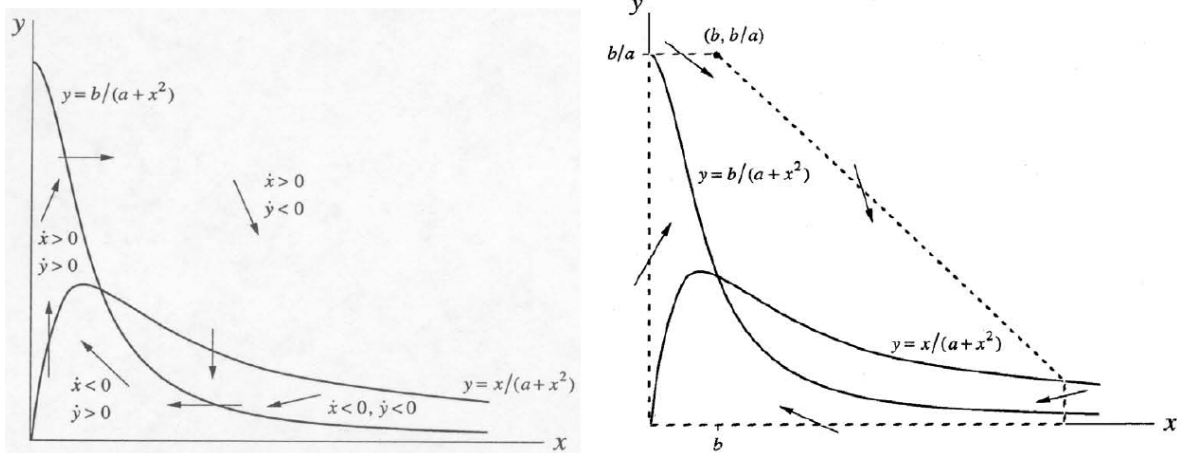
(b) Prove that this is a steady state.

$$(x, y) = \left(b, \frac{b}{a + b^2} \right).$$

(c) Show that there is a periodic orbit provided that $b^4 + (2a - 1)b^2 + (a + a^2) < 0$.

(d) Give an example of parameters a, b that satisfy this inequality.

9. We consider these equations:

$$dx/dt = y - f(x)$$
$$dy/dt = g(x) - y$$

where these are the plots of the two functions $y = f(x)$ and $y = g(x)$:
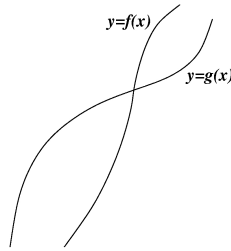
(a) Place arrows on nullclines, indicating directions, and put arrows in each connected component (regions delimited by nullclines), pointing NE, NW, etc.

(b) Let D be the region inside the dotted box. Show that D is a trapping region. (Place arrows showing which way the vector field points, on the boundary of D, that's all. Don't write too much.)
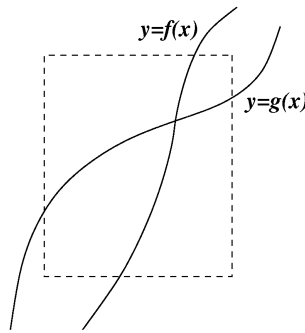
(c) What are the signs of the trace and the determinant of the Jacobian at the equilibrium shown in the picture?

(d) Can one apply the Poincaré-Bendixson Theorem to conclude that there must be a periodic orbit inside the dotted box?

(e) Can one apply the Bendixson criterion to conclude that there cannot be any periodic orbit inside the dotted box?

10. Prove that this system, where the constant a is nonzero:

$$\begin{aligned} dx_1/dt &= x_2 \\ dx_2/dt &= g(x_1) + ax_2 \end{aligned}$$

has no periodic orbit in $\mathbb{R}^2$.

11. Prove that this system, where the constants a, b, c satisfy that $c > a$:

$$\begin{aligned} dx_1/dt &= ax_1 - x_1 x_2 \\ dx_2/dt &= bx_1^2 - cx_2 \end{aligned}$$

has no periodic orbit inside the domain $\{x \in \mathbb{R}^2 \mid x_2 \geq 0\}$.

12. Let $B(x_1) = be^{-2\beta x_1}$. Show that the system

$$\begin{aligned} dx_1/dt &= B(x_1)\, x_2 \\ dx_2/dt &= B(x_1)\left(-ax_1 - bx_2 + \alpha x_1^2 + \beta x_2^2\right) \end{aligned}$$

has no periodic orbits in $\mathbb{R}^2$.

13. For the Van der Pol oscillator, when proving that the hexagonal region is a trapping region, we argued that we need not consider the following segments:

$$x = -3, \ -6 \le y \le 3 \qquad\qquad y = -6, \ -3 \le x \le 0 \qquad\qquad y = x - 6, \ 0 \le x \le 3$$

because of a symmetry argument. In this problem, you are asked to prove the property for these three segments, not using symmetries, but computing the dot product of an outward-facing normal and the vector field, and showing it is ≤ 0. (As explained in the text, don't worry about corners.)

Chapter 3

Deterministic PDE Models

3.1 Introduction to PDE models

Until now, we only considered functions of time (concentrations, populations, etc).
From now on, we consider functions that also depend on *space*.

A typical biological example of space-dependence would be the concentration of a morphogen as a function of space as well as time.

For example, this is a color-coded picture of a *Drosophila* embryo that has been stained for the protein products of genes *giant* (blue), *eve* (red), and *Kruppel* (other colors indicate areas where two or all genes are expressed):

One may also study space-dependence of a particular protein in a single cell. For example, this picture[1] shows the gradients of G-proteins in response to chemoattractant binding to receptors in the surface of *Dictyostelium discoideum* amoebas:
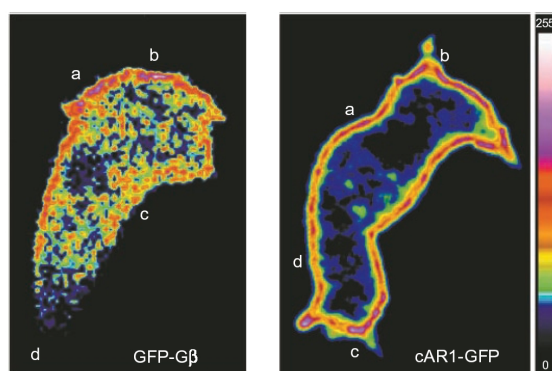
3.1.1 Densities

We write space variables as $x=(x_1, x_2, x_3)$ (or just (x, y) in dimension 2, or (x, y, z) in dimension 3).

---

[1]from Jin, Tian, Zhang, Ning, Long, Yu, Parent, Carole A., Devreotes, Peter N., "Localization of the G Protein Complex in Living Cells During Chemotaxis," *Science* 287(2000): 1034-1036.

We will work with *densities* "$c(x, t)$", which are understood intuitively in the following sense.

Suppose that we denote by $C(R, t)$ the amount of a type of particle (or number of individuals, mass of proteins of a certain type, etc.) in a region R of space, at time t.

Then, the density around point x, at time t, $c(x, t)$, is:

$$c(x, t) = \frac{C(\Delta R, t)}{\text{vol}(\Delta R)}$$

for "small" cubes ΔR around x, i.e. a "local average".

This means that $C(R, T) = \int\int\int_R c(x, t)\, dx$ for all regions R.
(A single or a double integral, if x is one- or two-dimensional, of course.)[2]

For now, we consider only scalar quantities $c(x, t)$; later we consider also vectors.

3.1.2 Reaction Term: Creation or Degradation Rate

We will assume that, at each point in space, there might take place a "reaction" that results in particles (individuals, proteins, bacteria, whatever) being created (or destroyed, depending on the sign).

This production (or decay) occurs at a certain rate "$\sigma(x, t)$" which, in general, depends on the location x and time t. (If there is no reaction, then $\sigma(x, t) = 0$.)

For scalar c, σ will typically be a formation or degradation rate.
More generally, if one considers vectors $c(x, t)$, with the coordinates of c representing for example the densities of different chemicals, then $\sigma(x, t)$ would represent the reactions among chemicals that happen to be in the same place at the same time.

The rate σ is a rate per unit volume per unit of time. That is, if $\Sigma(R, [a, b])$ is number of particles created (eliminated, if < 0) in a region R during time interval $[a, b]$, then the *average rate of growth* is:

$$\sigma(x, t) = \frac{\Sigma(\Delta R, [t, t + \Delta t])}{\text{vol}(\Delta R) \times \Delta t} \ ,$$

for "small" cubes ΔR around x and "small" time increments Δt. This means that

$$\Sigma(R, [a, b]) = \int_a^b \int\int\int_R \sigma(x, t)\, dx\, dt$$

for all regions R and time intervals $[a, b]$.

3.1.3 Conservation or Balance Principle

This is quite obvious:

increase (possibly negative) of quantity in a region = net creation + net influx.

Let us formalize this observation into an equation, studying first the one-dimensional case.

---

[2]In a more theoretical treatment of the subject, one would start with C, defined as a "measure" on subsets of $\mathbb{R}^3$, and the density c would be defined as a "derivative" of this measure C.

Suppose that R is a one-dimensional region along the x coordinate, defined by $x_1 \leq x \leq x_2$, and $c(x, t)$ and $\sigma(x, t)$ denote densities and reaction rates as a function of the scalar coordinate x.

Actually, it will be more convenient (and, in fact, is more realistic) to think of R as a three-dimensional volume, with a uniform cross-section in the y, z axes. Accordingly, we also think of the density $c(x, y, z, t) = c(x, t)$ and reaction rate $\sigma(x, y, z, t) = \sigma(x, t)$ as functions of a three-dimensional position (x, y, z), both uniform on each cross-section. We assume that nothing can "escape" through the y, z directions.

cross sectional area $= A$

positive x

R

x_1 x_2
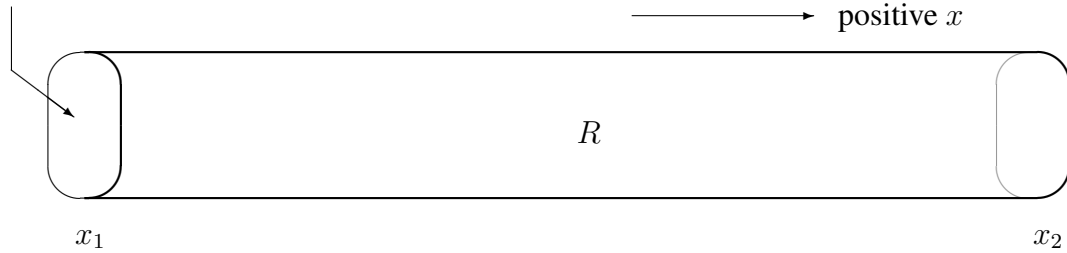
We need another important concept, the *flux*. It is defined as follows.
The flux at (x, t), written "$J(x, t)$", is the number of particles that cross a *unit area* perpendicular to x, in the positive direction, *per unit of time*.

x
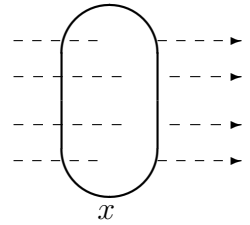
Therefore, the net flow through a cross-sectional area during a time interval $[a, b]$ is:

$$\int_a^b J(x, t) A \, dt \, .$$

We also need the following formulas, which follow from $\int_y \int_z = A$:

$$C(R, t) = \iiint_R c(\vec{x}, t) \, d\vec{x} = \int_{x_1}^{x_2} c(x, t) A \, dx \, ,$$

$$\Sigma(R, [a, b]) = \int_a^b \iiint_R \sigma(\vec{x}, t) \, d\vec{x} dt = \int_a^b \int_{x_1}^{x_2} \sigma(x, t) A \, dx dt \, .$$

We consider a segment $x \leq \xi \leq x + \Delta x$ and a time interval $[t, t + \Delta t]$.

ΔC

J_{in} J_{out}

$\Sigma = \int \sigma$

x $x + \Delta x$
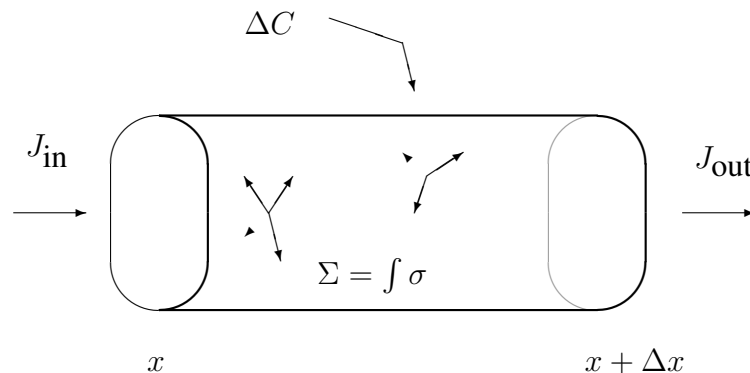
We have these equalities:

- net flow through cross-area at x: $\quad J_{\text{in}} \;=\; \displaystyle\int_t^{t+\Delta t} J(x,\tau)A\,d\tau$

- net flow through cross-area at $x+\Delta x$: $\quad J_{\text{out}} \;=\; \displaystyle\int_t^{t+\Delta t} J(x+\Delta x,\tau)A\,d\tau$

- net creation (elimination): $\quad \Sigma \;=\; \displaystyle\int_t^{t+\Delta t}\int_x^{x+\Delta x} \sigma(\xi,\tau)A\,d\xi\,d\tau$

- starting amount in segment: $\quad C_t \;=\; \displaystyle\int_x^{x+\Delta x} c(\xi,t)A\,d\xi$

- ending amount in segment: $\quad C_{t+\Delta t} \;=\; \displaystyle\int_x^{x+\Delta x} c(\xi,t+\Delta t)A\,d\xi.$

Finally, the change in total amount must balance-out:

$$C_{t+\Delta t} - C_t \;=\; \Delta C \;=\; J_{\text{in}} - J_{\text{out}} + \Sigma\,.$$

We have, putting it all together:

$$\int_x^{x+\Delta x} \left(c(\xi,t+\Delta t) - c(\xi,t)\right) A\,d\xi \;=\; \int_t^{t+\Delta t}\left(J(x,\tau) - J(x+\Delta x,\tau)\right) A\,d\tau + \int_t^{t+\Delta t}\int_x^{x+\Delta x} \sigma(\xi,\tau)A\,d\xi\,d\tau\,.$$

So, dividing by "$A\Delta t$", letting $\Delta t \to 0$, and applying the Fundamental Theorem of Calculus:

$$\int_x^{x+\Delta x} \frac{\partial c}{\partial t}(\xi,t)\,d\xi \;=\; J(x,t) - J(x+\Delta x,t) + \int_x^{x+\Delta x}\sigma(\xi,t)d\xi\,.$$

Finally, dividing by Δx, taking $\Delta x \to 0$, and once again using the FTC, we conclude:

$$\boxed{\frac{\partial c}{\partial t} \;=\; -\,\frac{\partial J}{\partial x} + \sigma}$$

This is the basic equation that we will use from now on.

We only treated the one-dimensional (i.e., uniform cross-section) case. However, the general case, when R is an arbitrary region in 3-space (or in 2-space) is totally analogous. One must define the flux $J(x,t)$ as a *vector* which indicates the maximal-flow direction at (x,t); its magnitude indicates the number of particles crossing, per unit time, a unit area perpendicular to J.

One derives, using Gauss' theorem, the following equation:

$$\boxed{\frac{\partial c}{\partial t} \;=\; -\,\operatorname{div} J + \sigma}$$

where the *divergence* of $J = (J_1, J_2, J_3)$ at $x = (x_1, x_2, x_3)$ is

$$\operatorname{div} J \;=\; \text{``}\nabla \cdot J\text{''} \;=\; \frac{\partial J_1}{\partial x_1} + \frac{\partial J_2}{\partial x_2} + \frac{\partial J_3}{\partial x_3}\,.$$

In the scalar case, $\operatorname{div} J$ is just $\frac{\partial J}{\partial x}$, of course.

Until now, everything was quite abstract. Now we specialize to very different types of fluxes.

3.1.4   Local fluxes: transport, chemotaxis

We first consider fluxes that arise from local effects, possibly infuenced by physical or chemical gradients.

3.1.5   Transport Equation

We start with the simplest type of equation, the *transport* (also known as the "convection" or the "advection" equation[3]).

We consider here *flux is due to transport*: a transporting tape as in an airport luggage pick-up, wind carrying particles, water carrying a dissolved substance, etc.

The main observation is that, in this case:

$$\text{flux} = \text{concentration} \times \text{velocity}$$

(depending on local conditions: x and t).

The following pictures may help in understanding why this is true.

flow direction; say constant speed

larger flux   smaller flux
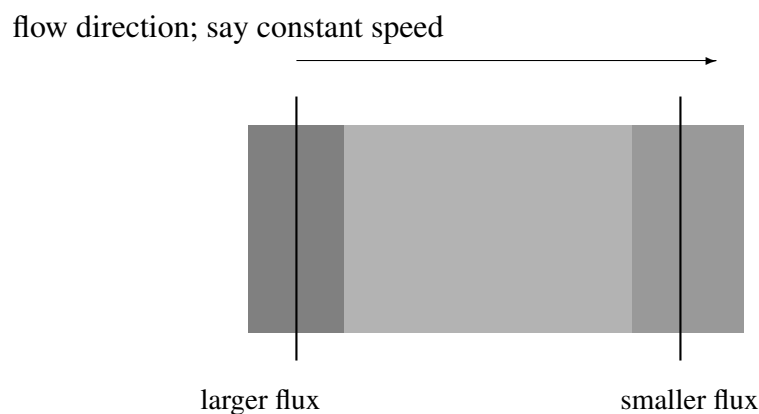
Let us zoom-in, approximating by a locally-constant density:

unit volumes, $c = 5$

unit area

flow at $v = 3$ units/sec

---
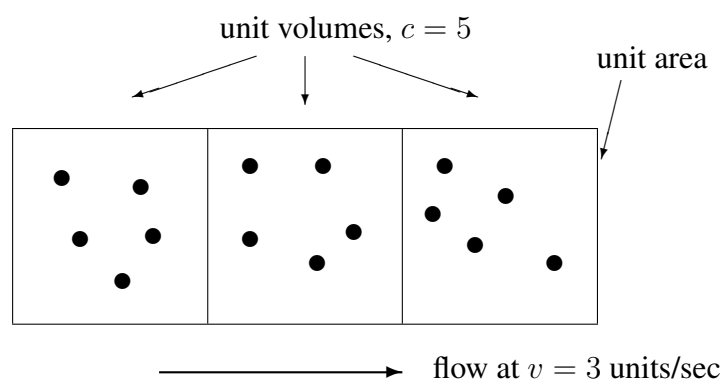
[3]In meteorology, convection and advection refer respectively to vertical and horizontal motion; the Latin origin is "advectio" = act of bringing.

Imagine a counter that "clicks" when each particle passes by the right endpoint. The total flux in one second is 15 particles. In other words, it equals cv. This will probably convince you of the following formula:

$$\boxed{J(x,t) \;=\; c(x,t)\,v(x,t)}$$

Since $\frac{\partial c}{\partial t} = -\operatorname{div} J + \sigma$, we obtain the *transport equation*:

$$\boxed{\frac{\partial c}{\partial t} \;=\; -\,\frac{\partial(cv)}{\partial x} \;+\; \sigma} \quad \text{or, equivalently:} \quad \boxed{\frac{\partial c}{\partial t} \;+\; \frac{\partial(cv)}{\partial x} \;=\; \sigma}$$

or more generally, in any dimension:

$$\boxed{\frac{\partial c}{\partial t} \;=\; -\,\operatorname{div}(cv) \;+\; \sigma} \quad \text{or, equivalently:} \quad \boxed{\frac{\partial c}{\partial t} \;+\; \operatorname{div}(cv) \;=\; \sigma}$$

This equation describes collective behavior, that of individual particles just "going with the flow".

Later, we will consider additional (and more interesting!) particle behavior, such as random movement, or movement in the direction of food. Typically, many such effects will be superimposed into the formula for J.

A special case is that of constant velocity $v(x,t) \equiv v$. For constant velocities, the above simplifies to:

$$\boxed{\frac{\partial c}{\partial t} \;=\; -\,v\frac{\partial c}{\partial x} \;+\; \sigma} \quad \text{or, equivalently:} \quad \boxed{\frac{\partial c}{\partial t} \;+\; v\frac{\partial c}{\partial x} \;=\; \sigma}$$

in dimension one, or more generally, in any dimension:

$$\boxed{\frac{\partial c}{\partial t} \;=\;=\; -v\operatorname{div}c \;+\; \sigma} \quad \text{or, equivalently:} \quad \boxed{\frac{\partial c}{\partial t} \;+\; \operatorname{div}c \;=\; \sigma}$$

Remark. If $\sigma = 0$, the equation becomes that of pure flow:

$$\frac{\partial c}{\partial t} + \operatorname{div}(cf) \;=\; 0$$

where are now writing "f" instead of "v" for the velocity, for reasons to be explained next. As before, let $c(x,t)$ denote the density of particles at location x and time t. The formula can be interpreted as follows. Particles move individually according to a differential equation $\frac{dx}{dt} = f(x,t)$. That is, when a particle is in location x at time t, its velocity should be $f(x,t)$. The equation then shows how the differential equation $\frac{dx}{dt} = f(x,t)$ for individual particles translates into a partial differential equation for densities. Seen in this way, the transport equation is sometimes called the *Liouville equation*. A special case is that in which $\operatorname{div}(f) = 0$, which is what happens in Hamiltonian mechanics. In that case, just as with constant velocity, we get the simplified equation $\frac{\partial c}{\partial t} + \sum_i \frac{\partial c}{\partial x_i} f_i$, where f_i is the ith coordinate of f. A probabilistic interpretation is also possible. Suppose that we think of single particles, whose initial conditions are distributed according to the density $c(x,0)$, and ask what is the probability density at time t. This density will be given by the solution of $\frac{\partial c}{\partial t} + \operatorname{div}(cf) = 0$, because we may think of an ensemble of particles, all evolving simultaneously. (It is implicit in this argument that particles are small enough that they never collide.)

3.1.6   Solution for Constant Velocity and Exponential Growth or Decay

Let us take the even more special case in which the reaction is linear: $\sigma = \lambda c$. This corresponds to a decay or growth that is proportional to the population (at a given time and place). The equation is:

$$\frac{\partial c}{\partial t} + v\frac{\partial c}{\partial x} = \lambda c$$

($\lambda > 0$ growth, $\lambda < 0$ decay).

Theorem: *Every solution (in dimension 1) of the above equation is of the form:*

$$c(x,t) = e^{\lambda t}f(x-vt)$$

for some (unspecified) differentiable single-variable function f.
Conversely, $e^{\lambda t}f(x-vt)$ is a solution, for any λ and f.

Notice that, in particular, when $t = 0$, we have that $c(x,0) = f(x)$. Therefore, the function f plays the role of an "initial condition" in time (but which depends, generally, on space).

The last part of the theorem is very easy to prove, as we only need to verify the PDE:

$$\left[\lambda e^{\lambda t}f(x-vt) - v e^{\lambda t}f'(x-vt)\right] + v e^{\lambda t}f'(x-vt) = \lambda e^{\lambda t}f(x-vt)\,.$$

Proving that the *only* solutions are these is a little more work:

we must prove that every solution of $\frac{\partial c}{\partial t} + v\frac{\partial c}{\partial x} = \lambda c$, where v and λ are given real constants), *must* have the form $c(x,t) = e^{\lambda t}f(x-vt)$, for some appropriate "$f$".

We start with the very special case $v = 0$. In this case, *for each fixed x, we have an ODE:* $\frac{\partial c}{\partial t} = \lambda c$.

Clearly, for each x, this ODE has the unique solution $c(x,t) = e^{\lambda t}c(x,0)$, so we can take $f(x)$ as the function $c(x,0)$.

The key step is to reduce the general case to this case, by "traveling" along the solution.
Formally, given a solution $c(x,t)$, we introduce a new variable $z = x - vt$, so that $x = z + vt$, and we define the auxiliary function $\alpha(z,t) := c(z+vt,t)$.

We note that $\frac{\partial\alpha}{\partial z}(z,t) = \frac{\partial c}{\partial x}(z+vt,t)$, but, more interestingly:

$$\frac{\partial\alpha}{\partial t}(z,t) = v\frac{\partial c}{\partial x}(z+vt,t) + \frac{\partial c}{\partial t}(z+vt,t)\,.$$

We now use the PDE $v\frac{\partial c}{\partial x} = \lambda c - \frac{\partial c}{\partial t}$ to get:

$$\frac{\partial\alpha}{\partial t}(z,t) = \left[\lambda c - \frac{\partial c}{\partial t}\right] + \frac{\partial c}{\partial t} = \lambda c(z+vt,t) = \lambda\alpha(z,t)\,.$$

We have thus reduced to the case $v = 0$ for α! So, $\alpha(z,t) = e^{\lambda t}\alpha(z,0)$. Therefore, substituting back:

$$c(x,t) = \alpha(x-vt,t) = e^{\lambda t}\alpha(x-vt,0)\,.$$

We conclude that

$$c(x,t) = e^{\lambda t}f(x-vt)$$

as claimed (writing $f(z) := \alpha(z,0)$).

Thus, all solutions are *traveling waves*, with decay or growth depending on the sign of λ.

These are typical figures, assuming that $v = 3$ and that $\lambda = 0$ and $\lambda < 0$ respectively (snapshots taken at $t = 0, 1, 2$):
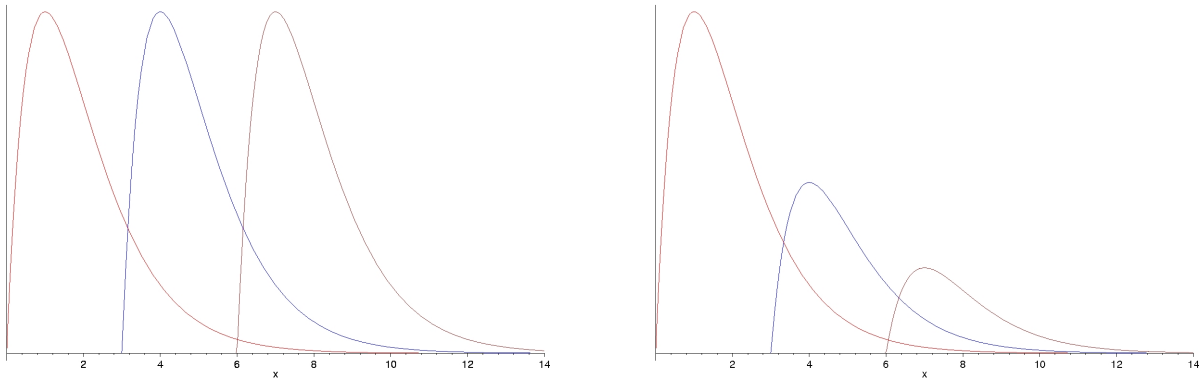
To determine uniquely $c(x, t) = e^{\lambda t} f(x - vt)$, need to know what the "initial condition f" is.

This could be done in various ways, for instance by specifying an initial distribution $c(x, 0)$, or by giving the values $c(x_0, t)$ at some point x_0.

Example: a nuclear plant is leaking radioactivity, and we measure a certain type of radioactive particle by a detector placed at $x = 0$. Let us assume that the signal detected is described by the following function:

$$h(t) = \begin{cases} 0 & t < 0 \\ \frac{1}{1+t} & t \geq 0 \end{cases} \,,$$

the wind blows eastward with constant velocity $v = 2$ m/s and particles decay with rate 3 s^{-1} ($\lambda = -3$). What is the solution $c(x, t)$?

We know that the solution is $c(x, t) = e^{-3t} f(x - 2t)$, but what is "$f$"?

We need to find f. Let us write the dummy-variable argument of f as "z" so as not to get confused with x and t. So we look for a formula for $f(z)$. *After we found $f(z)$*, we'll substitute $z = x - 2t$.

Since at position $x = 0$ we have that $c(0, t) = h(t)$, we know that $h(t) = c(0, t) = e^{-3t} f(-2t)$, which is to say, $f(-2t) = e^{3t} h(t)$.

We wanted $f(z)$, so we substitute $z = -2t$, and then obtain (since $t = -z/2$):

$$f(z) = e^{3(-z/2)} h(-z/2) \,.$$

To be more explicit, let us substitute the definition of h. Note that $t \geq 0$ is the same as $z \leq 0$. Therefore, we have:

$$f(z) = \begin{cases} \dfrac{e^{-3z/2}}{1 - z/2} & z \leq 0 \\ 0 & z > 0 \end{cases}$$

Finally, we conclude that the solution is:

$$c(x, t) = \begin{cases} \dfrac{e^{-3x/2}}{1 + t - x/2} & t \geq x/2 \\ 0 & t < x/2 \end{cases}$$

where we used the following facts: $z = x - 2t \leq 0$ is equivalent to $t \geq x/2$, $e^{-3t} e^{-(3/2)(x-2t)} = e^{-3x/2}$, and $1 - (x - 2t)/2 = 1 + t - x/2$.

We can now answer more questions. For instance: what is the concentration at position $x = 10$ and time $t = 6$? The answer is

$$c(10, 6) = \frac{e^{-15}}{2} \, .$$

3.1.7 Attraction, Chemotaxis

Chemotaxis is the term used to describe movement in response to chemoattractants or repellants, such as nutrients and poisons, respectively.

Perhaps the best-studied example of chemotaxis involves *E. coli* bacteria. In this course we will not study the behavior of individual bacteria, but will concentrate instead on the evolution equation for population density. However, it is worth digressing on the topic of individual bacteria, since it is so fascinating.

A Digression

E. coli bacteria are single-celled organisms, about 2 μm long, which possess up to six flagella for movement.

Chemotaxis in *E. coli* has been studied extensively. These bacteria can move in basically two modes: a "tumble" mode in which flagella turn clockwise and reorientation occurs, or a "run" mode in which flagella turn counterclockwise, forming a bundle which helps propel them forward.

Basically, when the cell senses a change in nutrient in a certain direction, it "runs" in that direction. When the sensed change is very low, a "tumble" mode is entered, with random reorientations, until a new direction is decided upon. One may view the bacterium as performing a stochastic gradient search in a nutrient-potential landscape. These are pictures of "runs" and "tumbles" performed by *E. coli*:
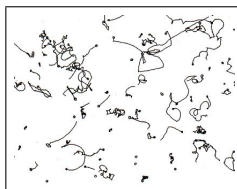
The runs are biased, drifting about 30 deg/s due to viscous drag and asymmetry. There is very little inertia (very low Reynolds number). The mean run interval is about 1 second and the mean tumble interval is about 1/10 sec.

The motors actuating the flagella are made up of several proteins. In the terms used by Harvard's Howard Berg[4], they constitute "a nanotechnologist's dream," consisting as they do of "engines, propellers, . . . , particle counters, rate meters, [and] gear boxes." These are an actual electron micrograph and a schematic diagram of the flagellar motor:
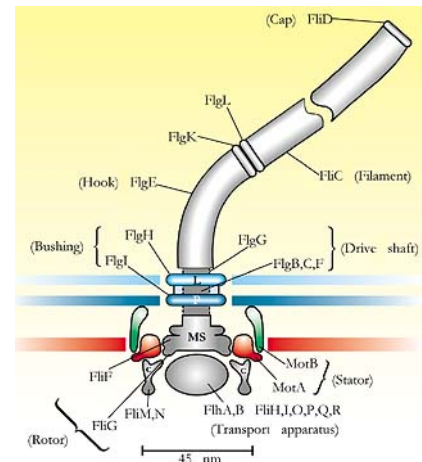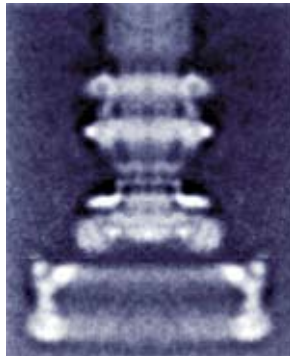
The signaling pathways involved in *E. coli* chemotaxis are fairly well understood. Aspartate or other nutrients bind to receptors, reducing the rate at which a protein called CheA ("Che" for "chemotaxis") phosphorylates another protein called CheY transforming it into CheY-P. A third protein, called CheZ, continuously reverses this phosphorylation; thus, when ligand is present, there is less CheY-P and more CheY. Normally, CheY-P binds to the base of the motor, helping clockwise movement and hence tumbling, so the lower concentration of CheY-P has the effect of less tumbling and more running (presumably, in the direction of the nutrient).

A separate feedback loop, which includes two other proteins, CheR and CheB, causes adaptation to constant nutrient concentrations, resulting in a resumption of tumbling and consequent re-orientation. In effect, the bacterium is able to take derivatives, as it were, and decide which way to go.
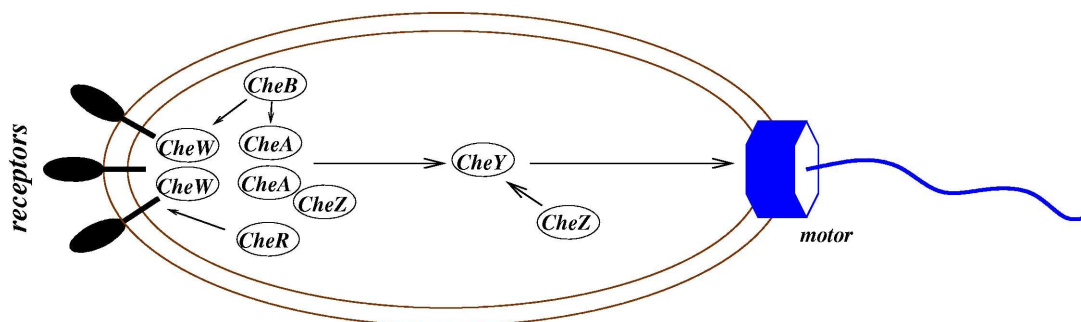
There are many papers (ask instructor for references if interested) describing biochemical models of how these proteins interact and mathematically analyzing the dynamics of the system.

---

[4]H. Berg, Motile behavior of bacteria, Physics Today, January 2000

Modeling how Densities Change due to Chemotaxis

Let us suppose given a function $V = V(x)$ which denotes the *concentration of a food source or chemical (or friends, or foes), at location*[5] x.

We think of V as a "potential" function, very much as with an electromagnetic or force field in physics.

The basic principle that we wish to model is: *the population is attracted toward places where V is larger*.

We often assume that either $V(x) \geq 0$ for all x or $V(x) \leq 0$ for all x.

We use the positive case to model attraction towards nutrient.

If V has negative values, then movement towards larger values of V means movement away from places where V is large in absolute value, that is to say, repulsion from such values, which might represent the locations of high concentrations of poisons or predator populations.

To be more precise: we will assume that individuals (in the population of which $c(x,t)$ measures the density) move at any given time *in the direction in which $V(x)$ increases the fastest when taking a small step*, and with a velocity that is proportional[6] to the perceived rate of change in magnitude of V.

We recall from multivariate calculus that $V(x+\Delta x) - V(x)$ maximized in the direction of its gradient.

The proof is as follows. We need to find a direction, i.e., unit vector "u", so that $V(x + hu) - V(x)$ is maximized, for any small stepsize h.

We take a linearization (Taylor expansion) for $h > 0$ small:

$$V(x + hu) - V(x) = [\nabla V(x) \cdot u] \, h + o(h).$$

This implies the following formula for the average change in V when taking a small step:

$$\frac{1}{h} \Delta V = \nabla V(x) \cdot u + O(h) \approx \nabla V(x) \cdot u$$

and therefore the maximum value is obtained precisely when the vector u is picked in the same direction as ∇V. Thus, the direction of movement is given by the gradient of V.

The magnitude of the vector $\frac{1}{h} \Delta V$ is the approximately $\nabla V(x)$. Thus, our assumptions give us that chemotaxis results in a velocity "$\alpha \, \nabla V(x)$" proportional to $\nabla V(x)$.

Since, in general, flux = density$\times$velocity, we conclude:

$$J(x,t) = \alpha \, c(x,t) \, \nabla V(x)$$

for some α, so that the obtained equation (ignoring reaction or transport effects) is:

$$\boxed{\frac{\partial c}{\partial t} = -\operatorname{div}(\alpha \, c \, \nabla V)} \quad \text{or, equivalently:} \quad \boxed{\frac{\partial c}{\partial t} + \operatorname{div}(\alpha \, c \, \nabla V) = 0}$$

and in particular, in the special case of dimension one:

$$\boxed{\frac{\partial c}{\partial t} = -\frac{\partial(\alpha \, c \, V')}{\partial x}} \quad \text{or, equivalently:} \quad \boxed{\frac{\partial c}{\partial t} + \frac{\partial(\alpha \, c \, V')}{\partial x} = 0}$$

---

[5]One could also consider time-varying functions $V(x,t)$. Time-varying V could help model a situation in which the "food" (e.g. a prey population) keeps moving.

[6]This is not always reasonable! Some other choices are: there is a maximum speed at which one can move, or movement is only possible at a fixed speed. See a homework problem.

and therefore, using the product rule for x-derivatives:

$$\frac{\partial c}{\partial t} = -\alpha \frac{\partial c}{\partial x} V' - \alpha c V''.$$

Of course, one can superimpose not only reactions but also different effects, such as transport, to this basic equation; the fluxes due to each effect add up to a total flux.

Example

Air flows (on a plane) Northward at 3 m/s, carrying bacteria. There is a food source as well, placed at $x = 1, y = 0$, which attracts according to the following potential:

$$V(x, y) = \frac{1}{(x-1)^2 + y^2 + 1}$$

(take $\alpha = 1$ and appropriate units).[7] The partial derivatives of V are:

$$\frac{\partial V}{\partial x} = -\frac{2x - 2}{((x-1)^2 + y^2 + 1)^2} \quad \text{and} \quad \frac{\partial V}{\partial y} = -2 \frac{y}{((x-1)^2 + y^2 + 1)^2}.$$

The differential equation is, then:

$$\frac{\partial c}{\partial t} = -\text{div}\,(c\nabla V) - \text{div}\left(\begin{pmatrix} 0 \\ 3 \end{pmatrix} c\right) = -\frac{\partial(c\frac{\partial V}{\partial x})}{\partial x} - \frac{\partial(c\frac{\partial V}{\partial y})}{\partial y} - 3\frac{\partial c}{\partial y}$$

or, expanding:

$$\frac{\partial c}{\partial t} = \frac{\partial c}{\partial x}\frac{(2x - 2)}{N^2} - 2c\frac{(2x-2)^2}{N^3} + 4\frac{c}{N^2} + 2\frac{\partial c}{\partial y}\frac{y}{N^2} - 8c\frac{y^2}{N^3} - 3\frac{\partial c}{\partial y}$$

where we wrote $N = (x-1)^2 + y^2 + 1$.

Some Intuition

Let us develop some intuition regarding the chemotaxis equation, at least in dimension one.

Suppose that we study what happens at a critical point of V. That is, we take a point for which $V'(x_0) = 0$. Suppose, further, that the concavity of V at that point is down: $V''(x_0) < 0$. Then, $\frac{\partial c}{\partial t}(x_0, t) > 0$, because:

$$\frac{\partial c}{\partial t}(x_0, t) = -\alpha\frac{\partial c}{\partial x}(x_0, t)\,V'(x_0) - \alpha c V''(x_0) = 0 - \alpha c V''(x_0) > 0.$$

In other words, the concentration at such a point increases in time. Why is this so, intuitively?

Answer: the conditions $V'(x_0) = 0$, $V''(x_0) < 0$ characterize a local maximum of V. Therefore, nearby particles (bacteria, whatever it is that we are studying) will move toward this point x_0, and the concentration there will increase in time.

---

[7]We assume that the food is not being carried by the wind, but stays fixed. (How would you model a situation where the food is also being carried by the wind?) Also, this model assumes that the amount of food is large enough that we need not worry about its decrease due to consumption by the bacteria. (How would you model food consumption?)

Conversely, if $V''(x_0) > 0$, then the formula shows that $\frac{\partial c}{\partial t}(x_0, t) < 0$, that is to say, the density decreases. To understand this intuitively, we can think as follows.

The point x_0 is a local minimum of V. Particles that start *exactly* at this point would not move, but any nearby particles will move "uphill" towards food. Thus, as nearby particles move away, the density at x_0, which is an average over small segments around x_0, indeed goes down.
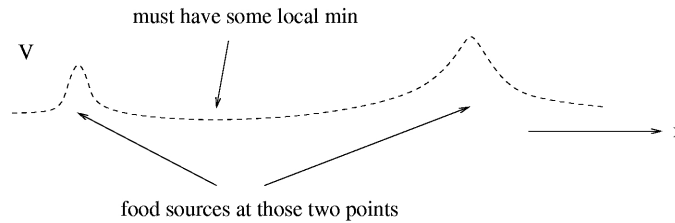
Next, let us analyze what happens when $V'(x_0) > 0$ and $V''(x_0) > 0$, under the additional assumption that $\frac{\partial c}{\partial x}(x_0, t) \approx 0$, that is, we assume that the density $c(x, t)$ is approximately constant around x_0. Then

$$\frac{\partial c}{\partial t}(x_0, t) = -\alpha \frac{\partial c}{\partial x}(x_0, t) V'(x_0) - \alpha c V''(x_0) \approx -\alpha c V''(x_0) < 0.$$

How can we interpret this inequality?

This picture of what the graph of V around x_0 looks like should help:
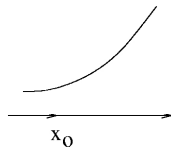
The derivative (gradient) of V is less to the left of x_0 than to the right of x_0, because $V'' > 0$ means that V' is increasing. So, the flux is less to the left of x_0 than to its right. This means that particles to the left of x_0 are arriving to the region around x_0 much slower than particles are leaving this region in the rightward direction. So the density at x_0 diminishes.

A homework problem asks you to analyze, in an analogous manner, these two cases:

(a) $V'(x_0) > 0, V''(x_0) < 0$
(b) $V'(x_0) < 0, V''(x_0) > 0$.

3.2 Non-local fluxes: diffusion

Diffusion is one of the fundamental processes by which "particles" (atoms, molecules, even bigger objects) move.

Fick's Law, proposed in 1855, and based upon experimental observations, postulated that diffusion is due to movement from higher to lower concentration regions. Mathematically:

$$J(x,t) \; \propto \; -\nabla c(x,t)$$

(we use "$\propto$" for "proportional").

This formula applies to movement of particles in a solution, where the proportionality constant will depend on the sizes of the molecules involved (solvent and solute) as well as temperature. It also applies in many other situations, such as for instance diffusion across membranes, in which case the constant depends on permeability and thickness as well.

The main physical explanation of diffusion is probabilistic, based on the thermal motion of individual particles due to the environment (e.g., molecules of solvent) constantly "kicking" the particles. "Brownian motion", named after the botanist Robert Brown, refers to such random thermal motion.

> One often finds the claim that Brown in his 1828 paper observed that pollen grains suspended in water move in a rapid but very irregular fashion.
>
> However, in *Nature*'s 10 March 2005 issue (see also errata in the 24 March issue), David Wilkinson states: "...several authors repeat the mistaken idea that the botanist Robert Brown observed the motion that now carries his name while watching the irregular motion of pollen grains in water. The microscopic particles involved in the characteristic jiggling dance Brown described were much smaller particles. I have regularly studied pollen grains in water suspension under a microscope without ever observing Brownian motion.
>
> From the title of Brown's 1828 paper "A Brief Account of Microscopical Observations ... on the Particles contained in the Pollen of Plants...", it is clear that he knew he was looking at smaller particles (which he estimated at about 1/500 of an inch in diameter) than the pollen grains.
>
> Having observed 'vivid motion' in these particles, he next wondered if they were alive, as they had come from a living plant. So he looked at particles from pollen collected from old herbarium sheets (and so presumably dead) but also found the motion. He then looked at powdered fossil plant material and finally inanimate material, which all showed similar motion.
>
> Brown's observations convinced him that life was not necessary for the movement of these microscopic particles."

The relation to Fick's Law was explained mathematically in Einstein's Ph.D. thesis (1905).[8]

When diffusion acts, and if there are no additional constraints, the eventual result is a homogeneous concentration over space. However, usually there are additional boundary conditions, creation and absorption rates, etc., which are superimposed on pure diffusion. This results in a "trade-off" between the "smoothing out" effects of diffusion and other influences, and the results can be very interesting.

We should also remark that diffusion is often used to model macroscopic situations analogous to movement of particles from high to low density regions. For example, a human population may shift towards areas with less density of population, because there is more free land to cultivate.

---

[8]A course project asks you to run a java applet simulation of Einstein's description of Brownian motion.

We have that $J(x,t) = -D \nabla c(x,t)$, for some constant D called the *diffusion coefficient*. Since, in general, $\frac{\partial c}{\partial t} = -\operatorname{div} J$, we conclude that:

$$\boxed{\frac{\partial c}{\partial t} = D \nabla^2 c}$$

where ∇^2 is the "Laplacian" (often "Δ") operator:

$$\frac{\partial c}{\partial t} = D \left(\frac{\partial^2 c}{\partial x_1^2} + \frac{\partial^2 c}{\partial x_2^2} + \frac{\partial^2 c}{\partial x_3^2} \right) .$$

The notation ∇^2 originates as follows: the divergence can be thought of as "dot product by ∇". So "$\nabla \cdot (\nabla c)$" is written as $\nabla^2 c$. This is the same as the "heat equation" in physics (which studies diffusion of heat).

Note that the equation is just:

$$\frac{\partial c}{\partial t} = D \frac{\partial^2 c}{\partial x^2}$$

in dimension one.

Let us consider the following very sketchy probabilistic intuition to justify why it is reasonable that the flux should be proportional to the gradient of the concentration, if particles move at random. Consider the following picture:
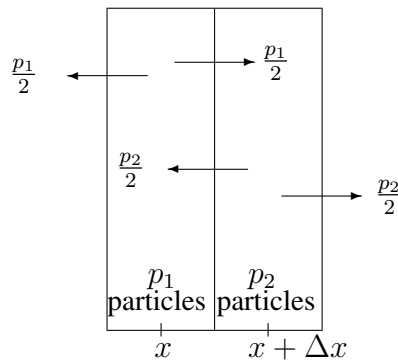
We assume that, in some small interval of time Δt, particles jump right or left with equal probabilities, so half of the p_1 particles in the first box move right, and the other half move left. Similarly for the p_2 particles in the second box. (We assume that the jumps are big enough that particles exit the box in which they started.)

The net number of particles (counting rightward as positive) through the segment shown in the middle is proportional to $\frac{p_1}{2} - \frac{p_2}{2}$, which is proportional roughly to $c(x,t) - c(x+\Delta x, t)$. This last difference, in turn, is proportional to $-\frac{\partial c}{\partial x}$.

This argument is not really correct, because we have said nothing about the velocity of the particles and how they relate to the scales of space and time. But it does intuitively help on seeing why the flux is proportional to the negative of the gradient of c.

A game can help understand. Suppose that students in a classroom all initially sit in the front rows, but then start to randomly (and repeatedly) change chairs, flipping coins to decide if to move backward (or forward if they had already moved back). Since no one is sitting in the back, initially there is a net flux towards the back. Even after a while, there will be still less students flipping coins in the back than in the front, so there are more possibilities of students moving backward than forward. Eventually, once that the students are well-distributed, about the same number will move forward as move backward: this is the equalizing effect of diffusion.

3.2.1 Time of Diffusion (in dimension 1)

It is often said that "diffusion results in movement proportional to $\sqrt{t}$". The following theorem gives one way to make that statement precise. A different interpretation is in the next section, and later, we will discuss a probabilistic interpretation and relations to random walks as well.

Theorem. Suppose that c satisfies diffusion equation

$$\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial x^2}.$$

Assume also that the following hold:

$$C = \int_{-\infty}^{+\infty} c(x,t)\,dx$$

is independent of t (constant population), and c is "small at infinity":

$$\text{for all } t \geq 0, \qquad \lim_{x \to \pm\infty} x^2\frac{\partial c}{\partial x}(x,t) = 0 \quad \text{and} \quad \lim_{x \to \pm\infty} xc(x,t) = 0\,.$$

Define, for each t, the following integral which measures how the density "spreads out":

$$\sigma^2(t) = \frac{1}{C}\int_{-\infty}^{+\infty} x^2 c(x,t)\,dx$$

(the second moment, which we assume is finite). Then:

$$\sigma^2(t) = 2Dt + \sigma^2(0)$$

for all t. In particular, if the initial (at time $t = 0$) population is concentrated near $x = 0$ (a "δ function"), then $\sigma^2(t) \approx 2Dt$.

Proof:

We use the diffusion PDE, and integrate by parts twice:

$$\frac{C}{D}\frac{d\sigma^2}{dt} = \frac{1}{D}\frac{\partial}{\partial t}\int_{-\infty}^{+\infty} x^2 c\,dx = \frac{1}{D}\int_{-\infty}^{+\infty} x^2\frac{\partial c}{\partial t}\,dx = \int_{-\infty}^{+\infty} x^2\frac{\partial^2 c}{\partial x^2}\,dx$$

$$= \left[x^2\frac{\partial c}{\partial x}\right]_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} 2x\frac{\partial c}{\partial x}\,dx$$

$$= -[2xc]_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} 2c\,dx = 2\int_{-\infty}^{+\infty} c(x,t)\,dx = 2C$$

Canceling C, we obtain:

$$\frac{d\sigma^2}{dt}(t) = 2D$$

and hence, integrating over t, we have, as wanted:

$$\sigma^2(t) = 2Dt + \sigma^2(0)\,.$$

If, in particular, particles start concentrated in a small interval around $x = 0$, we have that $c(x,0) = 0$ for all $|x| > \varepsilon$. Then, (with $c = c(x,0)$):

$$\int_{-\infty}^{+\infty} x^2 c\,dx = \int_{-\varepsilon}^{+\varepsilon} x^2 c\,dx \leq \varepsilon^2\int_{-\varepsilon}^{+\varepsilon} c\,dx = \varepsilon^2 C$$

so $\sigma^2(0) = \varepsilon^2 \approx 0$.

3.2.2 Another Interpretation of Diffusion Times (in dimension one)

There are many ways to state precisely what is meant by saying that diffusion takes time r^2 to move distance r. As diffusion is basically a model of a population of individuals which move randomly, one cannot talk about any particular particle, bacterium, etc. One must make a statement about the whole population. One explanation is in terms of the second moment of the density c, as done earlier. Another one is probabilistic, and one could also argue in terms of the Gaussian fundamental solution. We sketch another one next.

Suppose that we consider the diffusion equation $\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial x^2}$ for $x \in \mathbb{R}$, and an initial condition at $t = 0$ which is a step function, a uniform population density of one in the interval $(-\infty, 0]$ and zero for $x > 0$. It is quite intuitively clear that diffusion will result in population densities that look like the two subsequent figures, eventually converging to a constant value of 0.5.
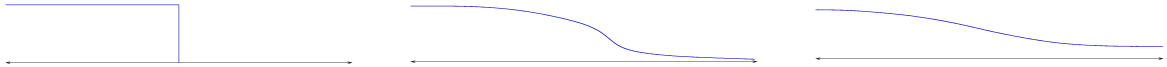
Consider, for any given coordinate point $p > 0$, a time $T = T(p)$ for which it is true that (let us say) $c(p, T) = 0.1$. It is intuitively clear (we will not prove it) that the function $T(p)$ is increasing on p: for those points p that are farther to the right, it will take longer for the graph to rise enough. So, $T(p)$ is uniquely defined for any given p. We sketch now a proof of the fact that $T(p)$ is proportional to p^2.

Suppose that $c(x, t)$ is a solution of the diffusion equation, and, for any given positive constant r, introduce the new function f defined by:

$$f(x, t) = c(rx, r^2 t).$$

Observe (chain rule) that $\frac{\partial f}{\partial t} = r^2 \frac{\partial c}{\partial t}$ and $\frac{\partial^2 f}{\partial x^2} = r^2 \frac{\partial^2 c}{\partial x^2}$. Therefore,

$$\frac{\partial f}{\partial t} - D\frac{\partial^2 f}{\partial x^2} = r^2\left(\frac{\partial c}{\partial t} - D\frac{\partial^2 c}{\partial x^2}\right) = 0.$$

In other words, the function f also satisfies the same equation. Moreover, c and f have the same initial condition: $f(x, 0) = c(rx, 0) = 1$ for $x \leq 0$ and $f(x, 0) = c(rx, 0) = 0$ for $x > 0$. Therefore f and c must be the same function.[9] In summary, for every positive number r, the following scaling law is true:

$$c(x, t) = c(rx, r^2 t) \quad \forall x, t.$$

For any $p > 0$, if we plug-in $r = p$, $x = 1$, and $t = T(p)/p^2$ in the above formula, we obtain that:

$$c(1, T(p)/p^2) = c(p.1, p^2.(T(p)/p^2)) = c(p, T(p)) = 0.1,$$

and therefore $T(1) = T(p)/p^2$, that is, $T(p) = \alpha p^2$ for some constant.

3.2.3 Separation of Variables

Let us try to find a solution of the diffusion equation, in dimension 1:

$$\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial x^2}$$

---

[9]Of course, uniqueness of solutions requires a proof. The fact that f and c satisfy the same "boundary conditions at infinity" is used in such a proof, which we omit here.

of the special form $c(x,t) = X(x)T(t)$.

Substituting into the PDE, we conclude that X, T must satisfy:

$$T'(t)X(x) \;=\; D\,T(t)X''(x)$$

(using primes for derivatives with respect to t and x), and this must hold for all t and x, or equivalently:

$$D\,\frac{X''(x)}{X(x)} \;=\; \frac{T'(t)}{T(t)} \quad \forall\, x, t \,.$$

Now define:

$$\lambda := \frac{T'(0)}{T(0)}$$

so:

$$D\frac{X''(x)}{X(x)} \;=\; \frac{T'(0)}{T(0)} \;=\; \lambda$$

for all x (since the above equality holds, in particular, at $t = 0$). Thus, we conclude, applying the equality yet again:

$$\boxed{D\,\frac{X''(x)}{X(x)} \;=\; \frac{T'(t)}{T(t)} \;=\; \lambda \quad \forall\, x, t}$$

for this fixed (and so far unknown) real number λ.

In other words, each of X and T satisfy an *ordinary* (and linear) differential equation, but the two equations share the same λ:

$$\begin{aligned}
X''(x) &= \lambda X(x) \\
T'(t) &= \lambda T(t) \,.
\end{aligned}$$

(We take $D=1$ for simplicity.) The second of these says that $T' = \lambda T$, i.e.

$$T(t) = e^{\lambda t}T(0)$$

and the first equation has the general solution (if $\lambda \neq 0$) $X(x) = ae^{\mu_1 x} + be^{\mu_2 x}$, where the μ_i's are the two square roots of λ, and a, b are arbitrary constants. As you saw in your diff equs course, when $\lambda < 0$, it is more user-friendly to write complex exponentials as trigonometric functions, which also has the advantage that a, b can then be taken as real numbers (especially useful since a and b are usually fit to initial conditions). In summary, for $\lambda > 0$ one has:

$$X(x) = ae^{\mu x} + be^{-\mu x}$$

(with $\mu = \sqrt{\lambda}$), while for $\lambda < 0$ one has:

$$X(x) = a\cos kx + b\sin kx$$

(with $k = \sqrt{-\lambda}$).

3.2.4   Examples of Separation of Variables

Suppose that a set of particles undergo diffusion (e.g., bacteria doing a purely random motion) inside a thin tube.

The tube is open at both ends, so part of the population is constantly being lost (the density of the organisms outside the tube is small enough that we may take it to be zero).

We model the tube in dimension 1, along the x axis, with endpoints at $x = 0$ and $x = \pi$:
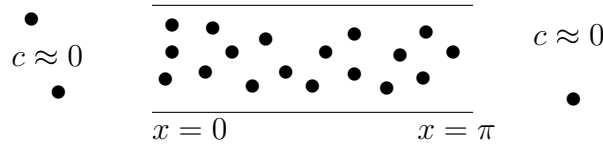
We model the problem by a diffusion (for simplicity, we again take $D{=}1$) with boundary conditions:

$$\frac{\partial c}{\partial t} = \frac{\partial^2 c}{\partial x^2}, \qquad c(0,t) = c(\pi,t) = 0.$$

Note that c identically zero is always a solution. Let's look for a bounded and nonzero solution.

Solution: we look for a $c(x,t)$ of the form $X(x)T(t)$. As we saw, if there is such a solution, then there is a number λ so that $X''(x) = \lambda X(x)$ and $T'(t) = \lambda T(t)$ for all x, t, so, in particular, $T(t) = e^{\lambda t}T(0)$. Since we were asked to obtain a *bounded* solution, the only possibility is $\lambda \leq 0$ (otherwise, $T(t) \to \infty$ as $t \to \infty$).

It cannot be that $\lambda = 0$. Indeed, if that were to be the case, then $X''(x) = 0$ implies that X is a line: $X(x) = ax + b$. But then, the boundary conditions $X(0)T(t) = 0$ and $X(\pi)T(t) = 0$ for all t imply that $ax + b = 0$ at $x = 0$ and $x = \pi$, giving $a = b = 0$, so $X \equiv 0$, but we are looking for a nonzero solution.

We write $\lambda = -k^2$, for some $k > 0$ and consider the general form of the X solution:

$$X(x) = a \sin kx + b \cos kx.$$

The boundary condition at $x = 0$ can be used to obtain more information:

$$X(0)T(t) = 0 \;\; \text{for all } t \;\; \Rightarrow\; X(0) = 0 \Rightarrow b = 0.$$

Therefore, $X(x) = a \sin kx$, and $a \neq 0$ (otherwise, $c \equiv 0$). Now using the second boundary condition:

$$X(\pi)T(t) = 0 \;\; \text{for all } t \;\; \Rightarrow\; X(\pi) = 0 \Rightarrow \sin k\pi = 0$$

Therefore, k *must be an integer* (nonzero, since otherwise $c \equiv 0$).

We conclude that any separated-form solution must have the form

$$c(x,t) = a\, e^{-k^2 t} \sin kx$$

for some nonzero integer k. One can easily check that, indeed, any such function is a solution (homework problem).

Moreover, since the diffusion equation is linear, any linear combination of solutions of this form is also a solution.

For example,

$$5e^{-9t}\sin 3x - 33e^{-16t}\sin 4x$$

is a solution of our problem.

In population problems, we cannot allow c to be negative. Solutions such as the above one are negative when the trigonometric functions have negative values, so they are not physically meaningful. However, we could modify the problem by assuming, for example, that the density outside the tube is equal to some constantly maintained value, let is say, $c = 100$. Then, the PDE becomes

$$\frac{\partial c}{\partial t} = \frac{\partial^2 c}{\partial x^2}, \quad c(0,t) = c(\pi,t) = 100\,.$$

Since the PDE is linear, and since $c(x,t) \equiv 100$ is a solution, the function $\widetilde{c}(x,t) = c(x,t) - 100$ is a solution of the homogeneous problem in which $c(0,t) = c(\pi,t) = 0$. Thus, $\widetilde{c}(x,t) = a\,e^{-k^2 t}\sin kx$ is a solution of the homogeneous problem, which means that

$$c(x,t) = 100 + \widetilde{c}(x,t) = 100 + a\,e^{-k^2 t}\sin kx$$

is a solution of the problem with boundary conditions $= 100$, for each a and each nonzero integer k. More generally, considering sums of the separated-form solutions, we have that if the sum of the coefficients "a" is less than 100, then we are guaranteed that the solution stays always nonnegative. For example,

$$100 + 5e^{-9t}\sin 3x - 33e^{-16t}\sin 4x$$

solves the equation with boundary conditions $c(0,t) = c(\pi,t) = 100$ *and* is always nonnegative.

Fitting Initial Conditions

Next let's add the requirement[10] that the *initial condition* must be:

$$c(x,0) = 3\sin 5x - 2\sin 8x\,.$$

Now, we know that any linear combination of the form

$$\sum_{k \text{ integer}} a_k e^{-k^2 t}\sin kx$$

solves the equation. Since the initial condition has the two frequencies $5, 8$, we should obviously try for a solution of the form:

$$c(x,t) = a_5 e^{-25t}\sin 5x + a_8 e^{-64t}\sin 8x\,.$$

We find the coefficients by plugging-in $t = 0$:

$$c(x,0) = a_5\sin 5x + a_8\sin 8x = 3\sin 5x - 2\sin 8x\,.$$

So we take $a_5 = 3$ and $a_8 = -2$; and thus obtain finally:

$$c(x,t) = 3e^{-25t}\sin 5x - 2e^{-64t}\sin 8x\,.$$

---

[10]For simplicity, we will take boundary conditions to be zero, even if this leads to physically meaningless negative solutions; as earlier, we can simply add a constant to make the problem more realistic.

One can prove, although we will not do so in this course, that this is the unique solution with the given boundary and initial conditions.

This works in exactly the same way whenever the initial condition is a finite sum $\sum_k a_k \sin kx$. Ignoring questions of convergence, the same idea even works for an infinite sum $\sum_{k=0}^{\infty} a_k \sin kx$. But what if initial condition is not a sum of sines? A beautiful area of mathematics, *Fourier analysis*, tells us that it is possible to write *any* function defined on an interval as an infinite sum of this form. This is analogous to the idea of writing any function of x (not just polynomials) as a sum of powers x^i. You saw such expansions (Taylor series) in a calculus course.

The theory of expansions into sines and cosines is more involved (convergence of the series must be interpreted in a very careful way), and we will not say anything more about that topic in this course.

Here are some pictures of approximations, though, for an interval of the form $[0, 2\pi]$. In each picture, we see a function together with various approximants consisting of sums of an increasing number of sinusoidal functions (red is constant; orange is $a_0 + a_1 \sin x$, etc).
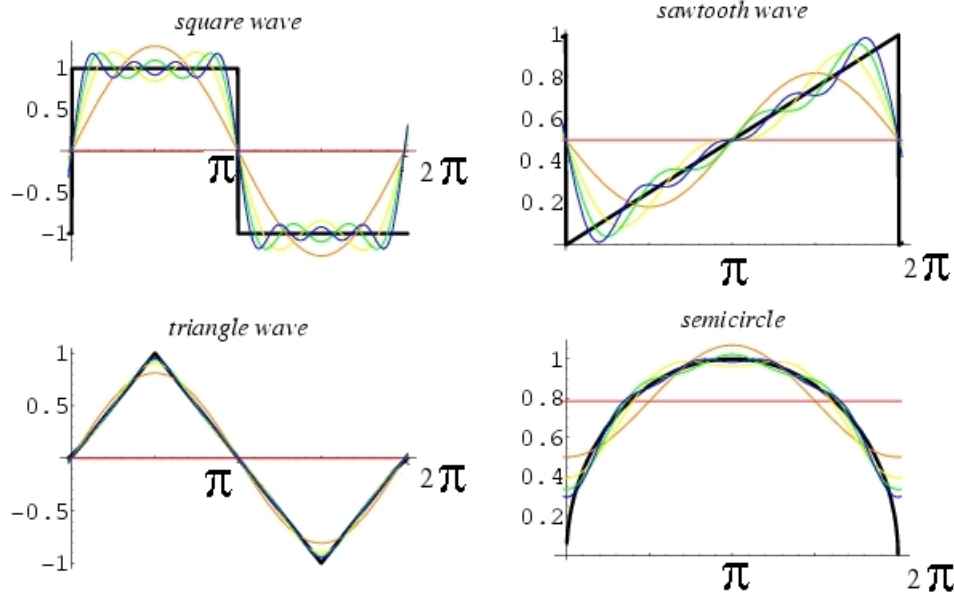
Another Example

Suppose now that, in addition to diffusion, there is a reaction. A population of bacteria grows exponentially inside the same thin tube that we considered earlier, still also moving at random.

We have this problem:

$$\frac{\partial c}{\partial t} \;=\; \frac{\partial^2 c}{\partial x^2} + \alpha c\,, \quad c(0,t) = c(\pi,t) = 0\,,$$

and look for nonzero solutions of the separated form $c(x,t) = X(x)T(t)$.

We follow the same idea as earlier:

$$X(x)T'(t) \;=\; X''(x)T(t) + \alpha X(x)T(t)$$

for all x, t, so there must exist some real number λ so that:

$$\frac{T'(t)}{T(t)} \;=\; \frac{X''(x)}{X(x)} + \alpha \;=\; \lambda\,.$$

This gives us the coupled equations:

$$\begin{aligned} T'(t) &= \lambda T(t) \\ X''(x) &= (\lambda - \alpha)X(x) \end{aligned}$$

with boundary conditions $X(0) = X(\pi) = 0$.

Suppose that $\lambda - \alpha \geq 0$. Then there is a real number μ such that $\mu^2 = \lambda - \alpha$ and X satisfies the equation $X'' = \mu^2 X$.

If $\mu = 0$, then the equation says that $X = a + bx$ for some a, b. But $X(0) = X(\pi) = 0$ would then imply $a = b = 0$, so $X \equiv 0$ and the solution is identically zero.

So let us assume that $\mu \neq 0$. Thus:

$$X = ae^{\mu x} + be^{-\mu x}$$

and, using the two boundary conditions, we have $a + b = ae^{\mu \pi} + be^{-\mu \pi} = 0$, or in matrix form:

$$\begin{pmatrix} 1 & 1 \\ e^{\mu \pi} & e^{-\mu \pi} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = 0 \,.$$

Since

$$\det \begin{pmatrix} 1 & 1 \\ e^{\mu \pi} & e^{-\mu \pi} \end{pmatrix} = e^{-\mu \pi} - e^{\mu \pi} = e^{-\mu \pi}(1 - e^{2\mu \pi}) \neq 0 \,,$$

we obtain that $a = b = 0$, again contradicting $X \not\equiv 0$. In summary, $\lambda - \alpha \geq 0$ leads to a contradiction, so $\lambda < \alpha$.

Let k be a real number such that $k^2 := \alpha - \lambda$. Then,

$$X'' + k^2 X = 0 \implies X(x) = a \sin kx + b \cos kx$$

and $X(0) = X(\pi) = 0$ implies that $b = 0$ and that k must be a nonzero integer.

For any give rate α, every separable solution is of form

$$a \, e^{(\alpha - k^2)t} \sin kx$$

with a nonzero integer k and some constant $a \neq 0$, and, conversely, every such function (or a linear combination thereof) is a solution (check!). If c represents a density of a population, a separable solution only makes sense if $k = 1$, since otherwise there will be negative values; however, sums of several such terms may well be positive. Note that if $\alpha > 1$ then there exists at elast one solution in which the population grows ($\alpha > k^2$, at least for $k = 1$).

3.2.5 No-flux Boundary Conditions

Suppose that the tube in the previous examples is closed at the end $x = L$ (a similar argument applies if it is closed at $x = 0$). We assume that, in that case, particles "bounce" at a "wall" placed at $x = L$.

One models this situation by a "no flux" or *Neumann* boundary condition $J(L, t) \equiv 0$, which, for the pure diffusion equation, is the same as $\frac{\partial c}{\partial x}(L, t) \equiv 0$.

One way to think of this is as follows. Imagine a narrow strip (of width ε) about the wall. For very small ε, most particles bounce back far into region, so the flux at $x = L - \varepsilon$ is ≈ 0.
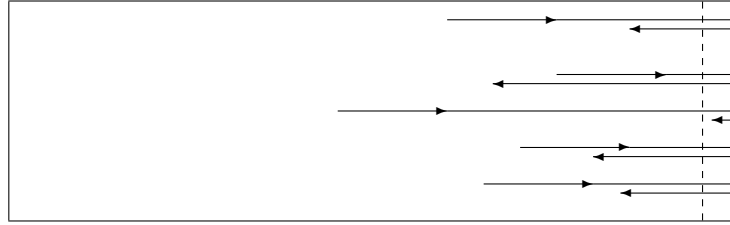
Another way to think of this is using the reflecting boundary method. We replace the wall by a "virtual wall" and look at equation in a larger region obtained by adding a mirror image of the original region. Every time that there is a bounce, we think of the particle as continuing to the mirror image section. Since everything is symmetric (we can start with a symmetric initial condition), clearly the net flow across this wall balances out, so even if individual particles would exit, *on the average* the same number leave as enter, and so the population density is exactly the same as if no particles would exit. As we just said, the flux at the wall must be zero, again explaining the boundary condition.
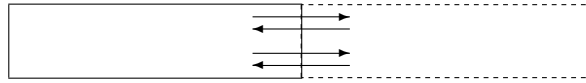
3.2.6   Probabilistic Interpretation

We make now some very informal and intuitive remarks.

In a population of indistinguishable particles (bacteria, etc.) undergoing random motion, we may track what happens to *each individual particle* (assumed small enough so that they don't collide with each other).

Since the particles are indistinguishable, one could imagine performing a huge number of one-particle experiments, and estimating the distribution of positions $x(t)$ by averaging over runs, instead of just performing one big experiment with many particles at once and measuring population density.

The probability of a single particle ending up, at time t, in a given region R, is proportional to how many particles there are in R, i.e. to Prob(particle in R) $\propto C(R, t) = \int_R c(x, t) \, dx$.

If we normalize to $C = 1$, we have that Prob(particle in R) = $\int_R c(x, t) \, dx$ (a triple integral, in 3 space).

Therefore, we may view $c(x, t)$ is the *probability density* of the random variable giving the position of an individual particle at time t (a *random walk*). In this interpretation, $\sigma^2(t)$ is then the *variance* of the random variable, and its standard deviation $\sigma(t)$ is proportional to $\sqrt{t}$ (a rough estimate on approximate distance traveled).

Specifically, for particles undergoing random motion with distribution c_0 (a "standard random walk"), the position has a Gaussian (normal) distribution.

For Gaussians, the mean distance from zero (up a to constant factor) coincides with the standard deviation:
$$E(|X|) \;=\; \frac{2}{\sigma\sqrt{2\pi}} \int_0^\infty x e^{-x^2/(2\sigma^2)} \, dx \;=\; \frac{\sigma}{\sqrt{\pi}}$$

(substitute $u = x/\sigma$), and similarly in any dimension for $E(\sqrt{x_1^2 + \ldots + x_d^2})$.

So we have that the *average displacement of a diffusing particle is proportional to $\sqrt{t}$*.

To put it in another way: *traveling average distance L requires time L^2.*

Since "life is motion" (basically by definition), this has fundamental implications for living organisms.

Diffusion is simple and energetically relatively "cheap": there is no need for building machinery for locomotion, etc., and no loss due to conversion to mechanical energy when running cellular motors and muscles.

At small scales, diffusion is very efficient (L^2 is tiny for small L), and hence it is a fast method for nutrients and signals to be carried along for *short* distances.

However, this is not the case for long distances (since L^2 is huge if L is large). Let's do some quick calculations.

Suppose that a particle travels by diffusion covering 10^{-6}m (= 1μm) in 10^{-3} seconds (a typical order of magnitude in a cell), Then, how much time is required to travel 1 meter?

Answer: since $x^2 = 2Dt$, we solve $(10^{-6})^2 = 2D10^{-3}$ to obtain $D = 10^{-9}/2$. So, $1 = 10^{-9}t$ means that $t = 10^9$ seconds, i.e. about 27 years!

Obviously, this is not a feasible way to move things along a large organism, or even a big cell (e.g., long neuron). That's one reason why there are circulatory systems, cell motors, microtubules, etc.

More on Random Walks

Let is develop a little more intuition on random walks. A discrete analog is as follows: suppose that a particle can move left or right with a unit displacement and equal probability, each step independent of the rest. What is the position after t steps? Let is check 4 steps, making a histogram:

ending	possible sequences	count	
-4	$-1-1-1-1$	1	x
-2	$-1-1-1+1,-1-1+1-1,...$	4	xxxx
0	$-1-1+1+1,-1+1+1-1,...$	6	xxxxxx
2	$1+1+1-1,1+1-1+1,...$	4	xxxx
4	$1+1+1+1$	1	x

The Central Limit Theorem tells us that the distribution (as $t \to \infty$ tends to be normal, with variance:

$$\sigma^2(t) = E(X_1 + \ldots + X_t)^2 = \sum\sum EX_iX_j = \sum EX_i^2 = \sigma^2 t$$

(since the steps are independent, $EX_iX_j = 0$ for $i \neq j$). We see then that $\sigma(t)$ is proportional to $\sqrt{t}$.

The theory of Brownian motion makes a similar analysis for continuous walks.

3.2.7   Another Diffusion Example: Population Growth

We consider now the equation

$$\frac{\partial c}{\partial t} = D\nabla^2 c + \alpha c \tag{3.1}$$

on the entire space (no boundary conditions).

This equation models a population which is diffusing and also reproducing at some rate α. It is an example of a *reaction-diffusion* equation, meaning that there is a reaction (in this case, $dc/dt = \alpha c$) taking place in addition to diffusion.

When there is no reaction ($\alpha = 0$), one can prove that the following "point-source" Gaussian formula:

$$p_0(x,t) = \frac{C}{\sqrt{4\pi Dt}} \exp\left(-\frac{x^2}{4Dt}\right) \tag{3.2}$$

is a solution in dimension 1, and a similar formula holds in higher dimensions (see problems). We use an integrating factor trick in order to reduce (3.1) to a pure diffusion equation. The trick is entirely analogous to what is done for solving the transport equation with a similar added reaction. We introduce the new dependent variable $p(x,t) := e^{-\alpha t} c(x,t)$. Then (homework problem), p satisfies the pure diffusion equation:

$$\frac{\partial p}{\partial t} = D\nabla^2 p\,.$$

Therefore, the solution for p is given by (3.2), and therefore

$$c(x,t) = \frac{C}{\sqrt{4\pi Dt}} \exp\left(\alpha t - \frac{x^2}{4Dt}\right)\,. \tag{3.3}$$

It follows that the equipopulation contours $c = $ constant have $x \approx \beta t$ for large t, where β is some positive constant. (A homework problem asks you to study this.)

This is noteworthy because, in contrast to the population dispersing a distance proportional to $\sqrt{t}$ (as with pure diffusion), the distance is, instead, proportional to t (which is much larger than $\sqrt{t}$). One intuitive explanation is that reproduction increases the gradient (the "populated" area has an even larger population) and hence the flux.

Similar results hold for the multivariate version, not just in dimension one.

Skellam[11] studied the spread of muskrats (*Ondatra zibethica*, a large aquatic rodent that originated in North America) in central Europe. Although common in Europe nowadays, it appears that their spread in Europe originated when a Bohemian farmer accidentally allowed several muskrats to escape, about 50 kilometers southwest of Prague. Diffusion with exponential growth followed.

The next two figures show the equipopulation contours and a plot of the square root of areas of spread versus time. (The square root of the area would be proportional to the distance from the source, if the equipopulation contours would have been perfect circles. Obviously, terrain conditions and locations of cities make these contours not be perfect circles.) Notice the match to the prediction of a linear dependence on time.

The third figure is an example[12] for the spread of Japanese beetles *Popillia japonica* in the Eastern United States, with invasion fronts shown.

---
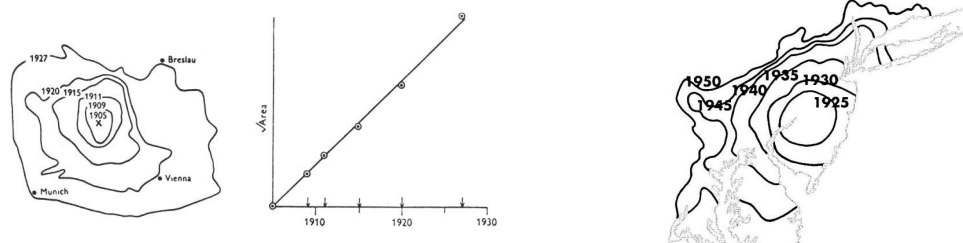
[11] J.G. Skellam, Random dispersal in theoretical populations, Biometrika 38: 196-218, 1951.

[12] from M.A. Lewis and S. Pacala, Modeling and analysis of stochastic invasion processes, J. Mathematical Biology 41, 387-429, 2000

Remark. Continuing on the topic of the Remark in page 164, suppose that each particle in a population evolves according to a differential equation $dx/dt = f(x,t) + w$, where "w" represents a "noise" effect which, in the absence of the f term, would make the particles undergo purely random motion, and the population density satisfies the diffusion equation with diffusion coefficient D. When both effects are superimposed, we obtain, for the density, an equation like $\partial c/\partial t = -\text{div}\,(cf) + D\nabla^2 c$. This is usually called a *Fokker-Planck* equation. (To be more precise, the Fokker-Planck equation describes a more general situation, in which the "noise" term affects the dynamics in a way that depends on the current value of x. We'll work out details in a future version of these notes.)

3.2.8 Systems of PDE's

Of course, one often must study *systems* of partial differential equations, not just single PDE's.

We just discuss one example, that of diffusion with growth and nutrient depletion, since the idea should be easy to understand. This example nicely connects with the material that we started the course with.

We assume that a population of bacteria, with density $n(x,t)$, move at random (diffusion), and in addition also reproduce with a rate $K(c(x,t))$ that depends on the local concentration $c(x,t)$ of nutrient.

The nutrient is depleted at a rate proportional to its use, and it itself diffuses. Finally, we assume that there is a linear death rate kn for the bacteria.

A model is:

$$
\begin{aligned}
\frac{\partial n}{\partial t} &= D_n \nabla^2 n + (K(c) - k)n \\
\frac{\partial c}{\partial t} &= D_c \nabla^2 c - \alpha K(c)n
\end{aligned}
$$

where D_n and D_c are diffusion constants. The function $K(c)$ could be, for example, a Michaelis-Menten rate $K(c) = \frac{k_{max}c}{k_n + c}$

You should ask yourself, as a homework problem, what the equations would be like if c were to denote, instead, a toxic agent, as well as formulate other variations of the idea.

Another example, related to this one, is that of chemotaxis with diffusion. We look at this example later, in the context of analyzing steady state solutions.

3.3 Steady-State Behavior of PDE's

In the study of ordinary differential equations (and systems) $\frac{dX}{dt} = F(X)$, a central role is played by steady states, that is, those states X for which $F(X) = 0$.

The vector field is only "interesting" near such states. One studies their stability, often using linearizations, in order to understand the behavior of the system under small perturbations from the steady state, and also as a way to gain insight into the global behavior of the system.

For a partial differential equation of the form $\frac{\partial c}{\partial t} = F(c, c_x, c_{xx}, \ldots)$, where c_x, etc., denote partial derivatives with respect to space variables, or more generally for systems of such equations, one may also look for steady states, and steady states also play an important role.

It is important to notice that, for PDE's, in general finding steady states involves not just solving an algebraic equation like $F(X) = 0$ in the ODE case, but a partial differential equation. This is because setting $F(c, c_x, c_{xx}, \ldots)$ to zero is a PDE on the space variables. The solution will generally be a function of x, not a constant. Still, the steady state equation is in general easier to solve; for one thing, there are less partial derivatives (no $\frac{\partial c}{\partial t}$).

For example, take the diffusion equation, which we write now as:

$$\frac{\partial c}{\partial t} = \mathcal{L}(c)$$

and where "$\mathcal{L}$" is the operator $\mathcal{L}(c) = \nabla^2 c$. A steady state is a function $c(x)$ that satisfies $\mathcal{L}(c) = 0$, that is,

$$\nabla^2 c = 0$$

(subject to whatever boundary conditions were imposed). This is the *Laplace equation*.

We note (but we have no time to cover in the course) that one may study stability for PDE's via "spectrum" (i.e., eigenvalue) techniques for a linearized system, just as done for ODE's.

To check if a steady state c_0 of $\frac{\partial c}{\partial t} = F(c)$ is stable, one linearizes at $c = c_0$, leading to $\frac{\partial c}{\partial t} = Ac$, and then studies the stability of the zero solution of $\frac{\partial c}{\partial t} = Ac$. To do that, in turn, one must find the eigenvalues and eigenvectors (now eigen-functions) of A (now an operator on functions, not a matrix), that is, solve

$$Ac = \lambda c$$

(and appropriate boundary conditions) for nonzero functions $c(x)$ and real numbers λ. There are many theorems in PDE theory that provide analogues to "stability of a linear PDE is equivalent to all eigenvalues having negative real part". To see why you may expect such theorems to be true, suppose that we have found a solution of $Ac = \lambda c$, for some $c \not\equiv 0$. Then, the function

$$\hat{c}(x, t) = e^{\lambda t} c(x)$$

solves the equation: $\frac{\partial \hat{c}}{\partial t} = A\hat{c}$. So, if for example, $\lambda > 0$, then $|\hat{c}(t, x)| \to \infty$ for those points x where $c(x) \neq 0$, as $t \to \infty$, and the zero solution is unstable. On the other hand, if $\lambda < 0$, then $\hat{c}(t, x) \to 0$.

For the Laplace equation, it is possible to prove that there are a countably infinite number of eigenvalues. If we write $\mathcal{L} = -\nabla^2 c$ (the negative is more usual in mathematics, for reasons that we will not explain here), then the eigenvalues of $\mathcal{L}$ form a sequence $0 < \lambda_0 < \lambda_1 < \ldots$, with $\lambda_n \to \infty$ as $n \to \infty$, when Dirichlet conditions (zero at boundary) are imposed, and $0 = \lambda_0 < \lambda_1 < \ldots$ when Neumann conditions (no-flux) are used. The eigenvectors that one obtains for domains that are

intervals are the trigonometric functions that we found when solving by separation of variables (the eigenvalue/eigenvector equation, for one space variable, is precisely $X''(x) + \lambda X = 0$).

In what follows, we just study steady states, and do not mention stability. (However, the steady states that we find turn out, most of them, to be stable.)

3.3.1   Steady State for Laplace Equation on Some Simple Domains

Many problems in biology (and other fields) involve the following situation. We have two regions, R and S, so that R "wraps around" S. A substance, such as a nutrient, is at a constant concentration, equal to c_0, on the exterior of R. It is also constant, equal to some other value c_S (typically, $c_S = 0$) in the region S. In between, the substance diffuses. See this figure:
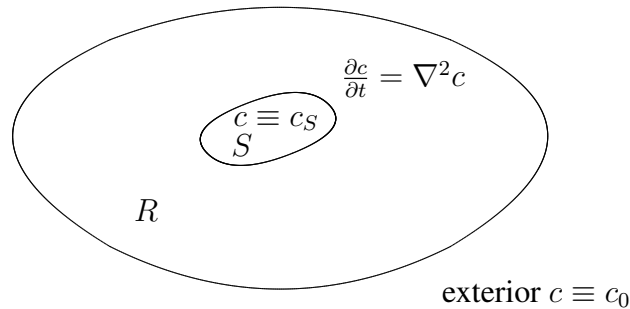
Examples abound in biology. For example, R might be a cell membrane, the exterior the extra-cellular environment, and S the cytoplasm.
In a different example, R might represent the cytoplasm and S the nucleus.
Yet another variation (which we mention later) is that in which the region R represents the immediate environment of a single-cell organism, and the region S is the organism itself.

In such examples, the external concentration is taken to be constant because one assumes that nutrients are so abundant that they are not affected by consumption. The concentration in S is also assumed constant, either because S is very large (this is reasonable if S would the cytoplasm and R the cell membrane) or because once nutrients enter S they get absorbed (combined with other substances) immediately (and so the concentration in S is $c_S = 0$).

Other examples typically modeled in this way include chemical transmitters at synapses, macrophages fighting infection at air sacs in lungs, and many others.

In this Section, we only study steady states, that is, we look for solutions of $\nabla^2 c = 0$ on R, with boundary conditions c_S and c_0.

Dimension 1

We start with the one-dimensional case, where S is the interval $[0, a]$, for some $a \geq 0$, and R is the interval $[a, L]$, for some $L > a$.

We view the space variable x appearing in the concentration $c(x, t)$ as one dimensional. However, one could also interpret this problem as follows: S and R are cylinders, there is no flux in the directions orthogonal to the x-axis, and we are only interested in solutions which are constant on cross-sections.

no flux

$$c(0,t) \equiv c_S \qquad \boxed{\quad c_t = D\nabla^2 c \quad} \qquad c(L,t) \equiv c_0$$

$$x = a \qquad\qquad \text{no flux} \qquad\qquad x = L$$
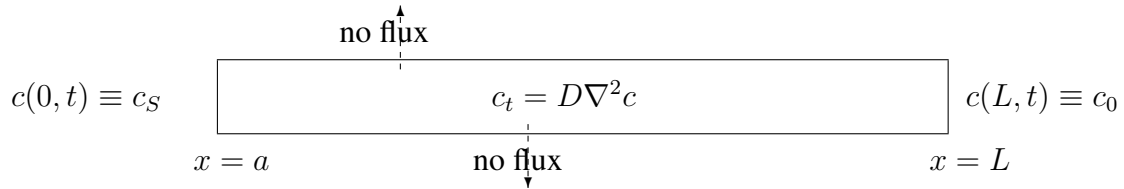
The steady-state problem is that of finding a function c of one variable satisfying the following ODE and boundary conditions:

$$D\frac{d^2 c}{dx^2} = 0, \quad c(a) = c_S, \ c(L) = c_0.$$

Since $c'' = 0$, $c(x)$ is linear, and fitting the boundary conditions gives the following unique solution:

$$c(x) = c_S + (c_0 - c_S)\frac{x - a}{L - a}.$$

Notice that, therefore, the gradient of c is $\frac{dc}{dx} = \frac{c_0 - c_S}{L - a}$.

Since, in general, the flux due to diffusion is $-D\nabla c$, we conclude that the flux is, in steady-state, the following constant:

$$J = -\frac{D}{L - a}(c_0 - c_S).$$

Suppose that $c_0 > c_S$. Then $J < 0$. In other words, an amount $\frac{D}{L-a}(c_0 - c_S)$ of nutrient transverses (from right to the left) the region $R = [a, L]$ per unit of time and per unit of cross-sectional area.

This formula gives an "Ohm's law for diffusion across a membrane" when we think of R as a cell membrane. To see this, we write the above equality in the following way:

$$c_S - c_0 = J\frac{L - a}{D}$$

which makes it entirely analogous to Ohm's law in electricity, $V = IR$. We interpret the potential difference V as the difference between inside and outside concentrations, the flux as current I, and the resistance of the circuit as the length divided by the diffusion coefficient. (Faster diffusion or shorter length results in less "resistance".)

Radially Symmetric Solutions in Dimensions 2 and 3

In dimension 2, we assume now that S is a disk of radius a and R is a washer with outside radius L. For simplicity, we take the concentration in S to be $c_S = 0$.
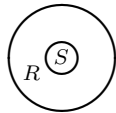
Since the boundary conditions are radially symmetric, we look for a radially symmetric solution, that is, a function c that depends only on the radius r.

Recalling the formula for the Laplacian as a function of polar coordinates, the diffusion PDE is:

$$\frac{\partial c}{\partial t} = \frac{D}{r}\frac{\partial}{\partial r}\left(r\frac{\partial c}{\partial r}\right), \quad c(a,t) = 0, \ c(L,t) = c_0.$$

Since we are looking only for a steady-state solution, we set the right-hand side to zero and look for $c = c(r)$ such that

$$(rc')' = 0 \quad c(a) = 0, \; c(L) = c_0,$$

where prime indicates derivative with respect to r.

A homework problem asks to show that the radially symmetric solution for the washer is:

$$c(r) = c_0 \frac{\ln(r/a)}{\ln(L/a)}.$$

Similarly, in dimension 3, taking S as a ball of radius a and R as the spherical shell with inside radius a and outside radius L, we have:

$$\frac{\partial c}{\partial t} = \frac{D}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial c}{\partial r} \right), \quad c(t, a) = 0, \; c(L, t) = c_0$$

The solution for the spherical shell is (homework problem):

$$c(r) = \frac{Lc_0}{L - a} \left(1 - \frac{a}{r} \right).$$

Notice the different forms of the solutions in dimensions 1, 2, and 3.

In the dimension 3 case, the derivative of c in the radial direction is, therefore:

$$c'(r) = \frac{Lc_0 a}{(L - a)r^2}.$$

We now specialize to the example in which the region R represents the environment surrounding a single-cell organism, the region S is the organism itself, and c models nutrient concentration.

We assume that the concentration of nutrient is constant far away from the organism, let us say farther than distance L, and $L \gg a$.

Then $c'(r) = \frac{c_0 a}{(1 - a/L)r^2} \approx \frac{c_0 a}{r^2}$.

In general, the steady-state flux due to diffusion, in the radial direction, is $-Dc'(r)$. In particular, on the boundary of S, where $r = a$, we have:

$$J = -\frac{Dc_0}{a}.$$

Thus $-J$ is the amount of nutrient that passes, in steady state, through each unit of area of the boundary, per unit of time. (The negative sign because the flow is toward the inside, i.e. toward smaller r, since $J < 0$.)

Since the boundary of S is a sphere of radius a, it has surface area $4\pi a^2$. Therefore, nutrients enter S at a rate of

$$\frac{Dc_0}{a} \times 4\pi a^2 = 4\pi Dc_0 a$$

per unit of time.

On the other hand, the metabolic need is roughly proportional to the volume of the organism. Thus, the amount of nutrients needed per unit of time is:

$$\frac{4}{3}\pi a^3 M\,,$$

where M is the metabolic rate per unit of volume per unit of time.

For the organism to survive, enough nutrients must pass through its boundary. If diffusion is the only mechanism for nutrients to come in, then survivability imposes the following constraint:

$$4\pi D c_0 a \ \geq\ \frac{4}{3}\pi a^3 M\,,$$

that is,

$$a \ \leq\ a_{\text{critical}} \ =\ \sqrt{\frac{3 D c_0}{M}}\,.$$

Phytoplankton[13] are free-floating aquatic organisms, and use bicarbonate ions (which enter by diffusion) as a source of carbon for photosynthesis, consuming one mole of bicarbonate per second per cubic meter of cell. The concentration of bicarbonate in seawater is about 1.5 moles per cubic meter, and $D \approx 1.5 \times 10^{-9} m^2 s^{-1}$. This gives

$$a_{\text{critical}} \ =\ \sqrt{3 \times 1.5 \times 10^{-9} \times 1.5\,m^2} \ \approx\ 82 \times 10^{-6}\,m = 82\,\mu m\,(\text{microns})\,.$$

This is, indeed, about the size of a typical "diatom" in the sea.

Larger organisms must use active transport mechanisms to ingest nutrients!

3.3.2   Steady States for a Diffusion/Chemotaxis Model

A very often used model that combines diffusion and chemotaxis is due to Keller and Segel. The model simply adds the diffusion and chemotaxis fluxes. In dimension 1, we have, then:

$$\frac{\partial c}{\partial t} \ =\ -\text{div}\,J \ =\ -\frac{\partial}{\partial x}\left(\alpha\,c\,V' - D\frac{\partial c}{\partial x}\right)\,.$$

We assume that the bacteria live on the one-dimensional interval $[0, L]$ and that no bacteria can enter or leave through the endpoints. That is, we have no flux on the boundary:[14]

$$J(0,t) = J(L,t) = 0 \quad \forall\,t\,.$$

Let us find the steady states.[15]

---

[13]We borrow this example from M. Denny and S. Gaines, *Chance in Biology*, Princeton University Press, 2000. The authors point out there that the metabolic need is more accurately proportional, for multicellular organisms, to $(\text{mass})^{3/4}$, but it is not so clear what the correct scaling law is for unicellular ones.

[14]Notice that this is not the same as asking that $\frac{\partial c}{\partial x}(0,t) = \frac{\partial c}{\partial x}(L,t) = 0$. The density might be constant near a boundary, but this does not mean that the population will not get redistributed, since there is also movement due to chemotaxis. Only for a pure diffusion, when $J = -D\frac{\partial c}{\partial x}$, is no-flux the same as $\frac{\partial c}{\partial x} = 0$.

[15]Note that, for a model in which there is only chemotaxis, there cannot be any stedy states, unless V was constant - why?

Setting $\frac{\partial c}{\partial t} = -\frac{\partial J}{\partial x} = 0$, and viewing now c as a function of x alone, and using primes for $\frac{d}{dx}$, gives:

$$J = \alpha\, c\, V' - D c' \;=\; J_0 \quad \text{(some constant)}.$$

Since $J_0 = 0$ (because J vanishes at the endpoints), we have that $(\ln c)' = c'/c = (\alpha V/D)'$, and therefore

$$c \;=\; k\, \exp(\alpha V/D)$$

for some constant. Thus, the steady state concentration is proportional to the exponential of the nutrient concentration, which is definitely not something that would be obvious.

For example, suppose that V is a concentration obtained from a steady-state gradient of a chemoattractant on $[0, L]$, where the concentration of V is zero at 0 and 1 at L. Then, $V(x) = x/L$ (prove!). It follows that, at steady state, $c(x) = ke^{x/L}$ (assuming for simplicity $D = 1$ and $\alpha = 1$).

3.3.3 Facilitated Diffusion

Let us now work out an example[16] involving a *system* of PDE's, diffusion, chemical reactions, and quasi-steady state approximations.

Myoglobin[17] is a protein that helps in the transport of oxygen in muscle fibers. The binding of oxygen to myoglobin results in oxymyoglobin, and this binding results in enhanced diffusion.

The facilitation of diffusion is somewhat counterintuitive, because the Mb molecule is much larger than oxygen (about 500 times larger), and so diffuses slower. A mathematical model helps in understanding what happens, and in quantifying the effect.

In the model, we take a muscle fibre to be one-dimensional, and no flux of Mb and MbO_2 in or out. (Only unbound oxygen can pass the boundaries.)

---

[16] Borrowing from J.P. Keener and J. Sneyd, Mathematical Physiology, Springer-Verlag New York, 1998.

[17] From Protein Data Bank, PDB, http://www.rcsb.org/pdb/molecules/mb3.html:

"myoglobin is where the science of protein structure really began...John Kendrew and his coworkers determined the atomic structure of myoglobin, laying the foundation for an era of biological understanding"

"The iron atom at the center of the heme group holds the oxygen molecule tightly. Compare the two pictures. The first shows only a set of thin tubes to represent the protein chain, and the oxygen is easily seen. But when all of the atoms in the protein are shown in the second picture, the oxygen disappears, buried inside the protein."

"So how does the oxygen get in and out, if it is totally surrounded by protein? In reality, myoglobin (and all other proteins) are constantly in motion, performing small flexing and breathing motions. Temporary openings constantly appear and disappear, allowing oxygen in and out. The structure in the PDB is merely one snapshot of the protein, caught when it is in a tightly-closed form"

$$s(0,t) \equiv s_0 \qquad \boxed{s = O_2,\ e = Mb,\ c = MbO_2} \qquad s(L,0) \equiv s_L \ll s_0$$
$$x = 0 \qquad\qquad\qquad\qquad\qquad\qquad x = L$$

The chemical reaction is just that of binding and unbinding:

$$O_2 + Mb \underset{k_-}{\overset{k_+}{\rightleftharpoons}} MbO_2$$
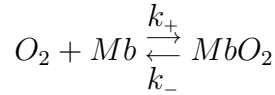
with equations:

$$\frac{\partial s}{\partial t} = D_s \frac{\partial^2 s}{\partial x^2} + k_- c - k_+ se$$
$$\frac{\partial e}{\partial t} = D_e \frac{\partial^2 e}{\partial x^2} + k_- c - k_+ se$$
$$\frac{\partial c}{\partial t} = D_c \frac{\partial^2 c}{\partial x^2} - k_- c + k_+ se,$$

where we assume that $D_e = D_c$ (since Mb and MbO_2 have comparable sizes). The boundary conditions are $\frac{\partial e}{\partial x} = \frac{\partial c}{\partial x} \equiv 0$ at $x = 0, L$, and $s(0) = s_0$, $s(L) = s_L$.

We next do a steady-state analysis of this problem, setting:

$$D_s s_{xx} + k_- c - k_+ se = 0$$
$$D_e e_{xx} + k_- c - k_+ se = 0$$
$$D_c c_{xx} - k_- c + k_+ se = 0$$

Since $D_e = D_c$, we have that $(e + c)_{xx} \equiv 0$.
So, $e + c$ is a linear function of x, whose derivative is zero at the boundaries (no flux).
Therefore, $e + c$ is constant, let us say equal to e_0.

On the other hand, adding the first and third equations gives us that

$$(D_s s_x + D_c c_x)_x = D_s s_{xx} + D_c c_{xx} = 0.$$

This means that $D_s s_x + D_c c_x$ is also constant:

$$D_s s_x + D_c c_x = -J.$$

Observe that J is the *the total flux of oxygen* (bound or not), since it is the sum of the fluxes $-D_s s_x$ of $s = O_2$ and $-D_c c_x$ of $c = MbO_2$.

Let $f(x) = D_s s(x) + D_c c(x)$. Since $f' = -J$, it follows that $f(0) - f(L) = JL$, which means:

$$J = \frac{D_s}{L}(s_0 - s_L) + \frac{D_c}{L}(c_0 - c_L)$$

(where one knows the oxygen concentrations s_0 and s_L, but not necessarily c_0 and c_L).

We will next do a quasi-steady state approximation, under the hypothesis that D_s is very small compared to the other numbers appearing in:

$$D_s s_{xx} + k_- c - k_+ s(e_0 - c) = 0$$

and this allows us to write[18]

$$c = e_0 \frac{s}{K + s}$$

where $K = k_-/k_+$. This allows us, in particular, to substitute c_0 in terms of s_0, and c_L in terms of s_L, in the above formula for the flux, obtaining:

$$J = \frac{D_s}{L}(s_0 - s_L) + \frac{D_c}{L}e_0\left(\frac{s_0}{K + s_0} - \frac{s_L}{K + s_L}\right).$$

This formula exhibits the flux as sum of the "Ohm's law" term plus plus a term that depends on diffusion constant D_c of myoglobin.

(Note that this second term, which quantifies the advantage of using myoglobin, is positive, since $s/(K + s)$ is increasing.)

With a little more work, which we omit here[19], one can solve for $c(x)$ and $s(x)$, using the quasi-steady state approximation. These are the graphs that one obtains, for the concentrations and fluxes respectively, of bound and free oxygen (note that the total flux J is constant, as already shown):
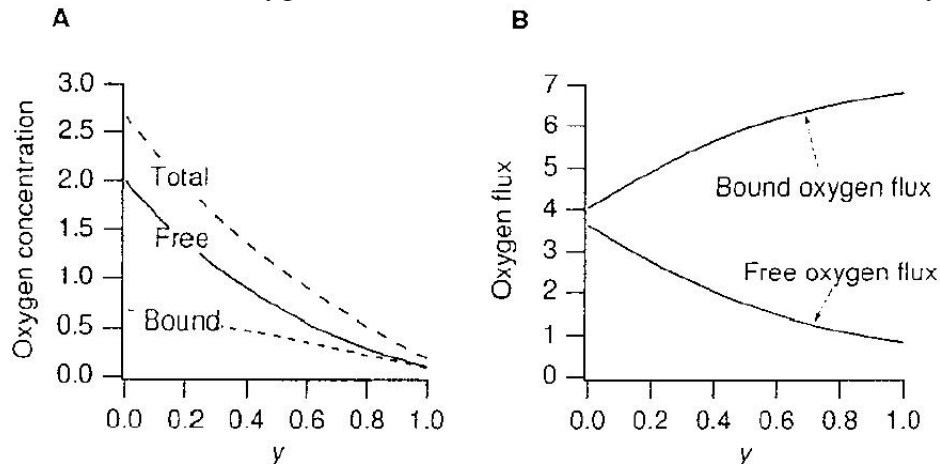
An intuition for why myoglobin helps is as follows. By binding to myoglobin, there is less free oxygen near the left endpoint. As the boundary conditions say that the concentration is s_0 outside, there is more flow into the cell (diffusion tends to equalize). Similarly, at the other end, the opposite happens, and more flows out.

3.3.4 Density-Dependent Dispersal

Here is yet another example[20] of modeling with a system of PDE's and steady-state calculations.

Suppose that the flux is proportional to $-c\nabla c$, not to $-\nabla c$ as with diffusion: a transport-like equation, where the velocity is determined by the gradient. In the scalar case, this would mean that the flux is proportional to $-cc_x$, which is the derivative of $-c^2$. Such a situation would occur if, for instance, overcrowding encourages more movement.

---

[18]Changing variables $\sigma = (k_+/k_-)s$, $u = c/e_0$, and $y = x/L$, one obtains $\varepsilon\sigma_{yy} = \sigma(1 - u) - u$, $\varepsilon = D_s/(e_0k_+L^2)$. A typical value of ε is estimated to be $\varepsilon \approx 10^{-7}$. This says that $\sigma(1 - u) - u \approx 0$, and from here one can solve for u as a function of σ, or equivalently, c as a function of s.

[19]see the Keener-Sneyd book for details

[20]from Keshet's book

So we have

$$\frac{\partial u}{\partial t} = -\text{div}\,(-\alpha u \nabla u)$$

and, in particular, in dimension 1:

$$\frac{\partial u}{\partial t} = \alpha \frac{\partial}{\partial x}\left(u\frac{\partial u}{\partial x}\right)$$

Let us look for steady states: $u = u(x)$ solving (with $u' = \frac{\partial u}{\partial x}$):

$$(uu')' = 0\,.$$

This means that $(u^2)'' = 0$, i.e. $u(x)^2 = ax + b$, or $u(x) = \sqrt{ax+b}$ for some constants a, b. Suppose that we also impose boundary conditions $u(0) = 1$ and $u(1) = 2$. Then, $u(x) = \sqrt{3x+1}$. The plot of $u(x)$ clearly shows that the total amount of individuals (the integral $\int_0^1 u(x)dx$) is larger than it would be if pure diffusion would occur, in which case $u(x) = x + 1$ (why?).

To make the problem a little more interesting, let us now assume that there are two interacting populations, with densities u and v respectively, and each moves with a velocity that is proportional to the gradient of the *total* population $u + v$.

We obtain these equations:

$$\frac{\partial u}{\partial t} = -\text{div}\,(-\alpha u \nabla(u+v))$$
$$\frac{\partial v}{\partial t} = -\text{div}\,(-\beta v \nabla(u+v))$$

and, in particular, in dimension 1:

$$\frac{\partial u}{\partial t} = \alpha \frac{\partial}{\partial x}\left(u\frac{\partial(u+v)}{\partial x}\right)$$
$$\frac{\partial v}{\partial t} = \beta \frac{\partial}{\partial x}\left(v\frac{\partial(u+v)}{\partial x}\right)\,.$$

Let us look for steady states: $u = u(x)$ and $v = v(x)$ solving (with $u' = \frac{\partial u}{\partial x}$, $v' = \frac{\partial v}{\partial x}$):

$$(u(u+v)')' = (v(u+v)')' = 0\,.$$

There must exist constants c_1, c_2 so that:

$$u(u+v)' = c_1\,,\quad v(u+v)' = c_2\,.$$

We study three separate cases:
(1) $c_1 = c_2 = 0$
(2) $c_2 \neq 0$ and $c_1 = 0$,
(3) $c_1 c_2 \neq 0$
(the case $c_1 \neq 0$ and $c_2 = 0$ is similar to (2)).

Case (1):
here $[(u+v)^2]' = 2(u+v)(u+v)' = 2u(u+v)' + 2v(u+v)' = 0$, so $u + v$ is constant. That's the best that we can say.

Case (2):

$$c_2 \neq 0 \;\Rightarrow\; v(x) \neq 0, \, (u+v)'(x) \neq 0 \; \forall \, x \,.$$

Also,

$$c_1 = 0 \;\Rightarrow\; u \equiv 0 \;\Rightarrow\; vv' \equiv c_2 \;\Rightarrow\; (v^2)' \equiv 2c_2$$

implies $v^2 = 2c_2 x + K$ for some constant K, so (taking the positive square root, because $v \geq 0$, being a population):

$$v \;=\; \sqrt{2c_2 x + K} \,, \quad u \equiv 0 \,.$$

Case (3):
Necessarily $u(x) \neq 0$ and $v(x) \neq 0$ for all x, so can divide and obtain:

$$(u+v)' = \frac{c_1}{u} = \frac{c_2}{v} \,.$$

Hence $u = (c_1/c_2)v$ can be substituted into $u' + v' = \frac{c_2}{v}$ to obtain $(1 + c_1/c_2)v' = c_2/v$, i.e. $vv' = c_2/(1 + c_1/c_2)$, or $(v^2)' = 2c_2/(1 + c_1/c_2)$, from which:

$$v^2(x) = \frac{2c_2 x}{1 + c_1/c_2} + K$$

for some K, and so:

$$v(x) = \left(\frac{2c_2 x}{1 + c_2/c_1} + K \right)^{1/2} \,.$$

Since $u = (c_1/c_2)v$,

$$u(x) = \left(\frac{2c_1 x}{1 + c_1/c_2} + K c_1^2/c_2^2 \right)^{1/2} \,.$$

3.4 Traveling Wave Solutions of Reaction-Diffusion Systems

It is rather interesting that reaction-diffusion systems can exhibit traveling-wave behavior. Examples arise from systems exhibiting bistability, such as the developmental biology examples considered earlier, or, in a more complicated system form, for species competition.

The reason that this is surprising is that diffusion times tend to scale like the square root of distance, not linearly. (But we have seen a similar phenomenon when discussing diffusion with exponential growth.)

We illustrate with a simple example, the following equation:

$$\frac{\partial V}{\partial t} = \frac{\partial^2 V}{\partial x^2} + f(V)$$

where f is a function that has zeroes at $0, \alpha, 1$, $\alpha < 1/2$, and satisfies:

$$f'(0) < 0, \ \ f'(1) < 0, \ \ f'(\alpha) > 0$$
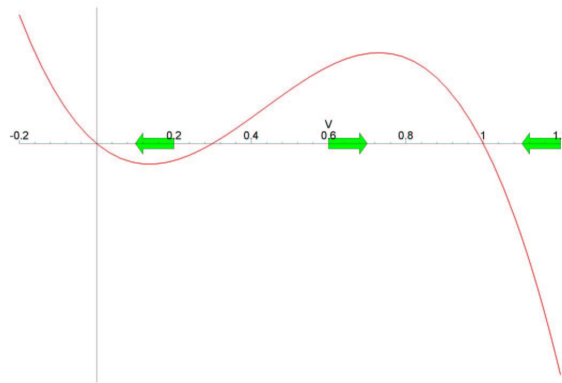
so that the differential equation $dV/dt = f(V)$ by itself, without diffusion, would be a bistable system.[21]

We would like to know if there's any solution that looks like a "traveling front" moving to the left (we could also ask about right-moving solutions, of course).
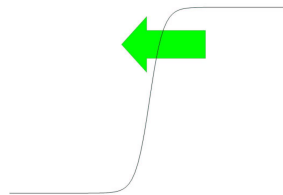
In other words, we look for $V(x, t)$ such that, for some "waveform" U that "travels" at some speed c, V can be written as a translation of U by ct:

$$V(x, t) = U(x + ct).$$

In accordance with the above picture, we also want that these four conditions hold:

$$V(-\infty, t) = 0, \ V(+\infty, t) = 1, \ V_x(-\infty, t) = 0, \ V_x(+\infty, t) = 0.$$

---

[21]Another classical example is that in which f represents logistic growth. That is the Fisher equation, which is used in genetics to model the spread in a population of a given allele.

The key step is to realize that the PDE for V induces an ordinary differential equation for the waveform U, and that these boundary conditions constrain what U and the speed c can be.

To get an equation for U, we plug-in $V(x,t) = U(x+ct)$ into $V_t = V_{xx} + f(V)$, obtaining:

$$cU' = U'' + f(U)$$

where "'" indicates derivative with respect to the argument of U, which we write as ξ. Furthermore, $V(-\infty, t) = 0$, $V(+\infty, t) = 1$, $V_x(-\infty, t) = 0$, $V_x(+\infty, t) = 0$ translate into:

$$U(-\infty) = 0\,,\ \ U(+\infty) = 1\,,\ \ U'(-\infty) = 0\,,\ \ U'(+\infty) = 0\,.$$

Since U satisfies a second order differential equation, we may introduce $W = U'$ and see U as the first coordinate in a system of two first-order differential equations:

$$\begin{aligned} U' &= W \\ W' &= -f(U) + cW\,. \end{aligned}$$

The steady states satisfy $W = 0$ and $f(U) = 0$, so they are $(0,0)$, $(1,0)$, as well as $(\alpha, 0)$. The Jacobian is

$$J = \begin{pmatrix} 0 & 1 \\ -f' & c \end{pmatrix}$$

and has determinant $f' < 0$ at the first two steady states, so they are both saddles. The conditions on U translate into the requirements that:

$$(U,W) \to (0,0) \text{ as } \xi \to -\infty \text{ and } (U,W) \to (1,0) \text{ as } \xi \to \infty$$

for the function $U(\xi)$ and its derivative, seen as a solution of this system of two ODE's. (Note that "ξ" is now "time".) In dynamical systems language, we need to show the existence of an "heteroclinic connection" between these two saddles. One first proves that, for $c \approx 0$ and $c \gg 1$, there result trajectories that "undershoot" or "overshoot" the desired connection, so, by a continuity argument (similar to the intermediate value theorem), there must be some value c for which the connection exactly happens. Details are given in many mathematical biology books.
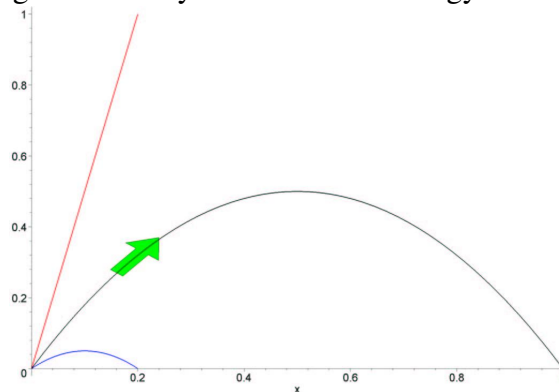
The theory can be developed quite generally, but here we'll only study in detail this very special case:

$$f(V) = -A^2 V(V - \alpha)(V - 1)$$

which is easy to treat with explicit formulas.

Since U will satisfy $U' = 0$ when $U = 0, 1$, we *guess* the functional relation:

$$U'(\xi) \;=\; BU(\xi)\,(1 - U(\xi))$$

(note that we are looking for a U satisfying $0 \le U \le 1$, so $1 - U \ge 0$). We write "ξ" for the argument of U so as to not confuse it with x.

We substitute $U' = BU(1 - U)$ and (taking derivatives of this expression)

$$U'' = B^2 U(1 - U)(1 - 2U)$$

into the differential equation $cU' = U'' + A^2 U(U - \alpha)(U - 1)$, and cancel $U(U - 1)$, obtaining (the calculation is given as a homework problem):

$$B^2(2U - 1) + cB - A^2(U - \alpha) \;=\; 0\,.$$

Since U is not constant (because $U(-\infty) = 0$ and $U(+\infty) = 1$), this means that we can compare coefficients of U in this expression, and conclude: that $2B^2 - A^2 = 0$ and $-B^2 + cB + \alpha A^2 = 0$. Therefore:

$$B = A/\sqrt{2}, \quad c = \frac{(1 - 2\alpha)A}{\sqrt{2}}\,.$$

Substituting back into the differential equation for U, we have:

$$U' \;=\; BU(1 - U) \;=\; \frac{A}{\sqrt{2}}U(1 - U)\,,$$

an ODE that now does not involve the unknown B. We solve this ODE by separation of variables and partial fractions, using for example $U(0) = 1/2$ as an initial condition, getting:

$$U(\xi) \;=\; \frac{1}{2}\left[1 + \tanh\left(\frac{A}{2\sqrt{2}}\xi\right)\right]$$

(a homework problem asks to verify that this is a solution). Finally, since $V(x, t) = U(x + ct)$, we conclude that:

$$V(x, t) \;=\; \frac{1}{2}\left[1 + \tanh\left(\frac{A}{2\sqrt{2}}(x + ct)\right)\right]$$

where $c = \frac{(1-2\alpha)A}{\sqrt{2}}$.

Observe that the speed c was uniquely determined; it will be larger if $\alpha \approx 0$, or if the reaction is stronger (larger A). This is not surprising! (Why?)

3.5 Problems for PDE chapter

Problems PDE1: Transport

1. Suppose that $c(x, t)$ is the density of bacteria (in one space dimension) being carried east by a wind blowing at 1 mph. The bacteria double every 1 hour. Suppose that we know that $c(1, t) = t$ for all t. Derive a formula for $c(x, t)$ for all x, t, by using the general form "$e^{\lambda t} f(x - vt)$" and determining the constant λ and the function f from the given information. You should end up with this answer:

$$(1 - x + t) e^{t \ln 2} e^{-(\ln 2)(1 - x + t)}$$

but *do not* just plug-in this expression to verify that it is a solution, since the whole point of the problem is for you to be able to work out the problem without knowing the solution!

2. Suppose that a population, with density $c(x, t)$ (in one dimension), is being transported with velocity $v = 5x$. That is to say, the velocity is not constant, but it depends on the position x. (This could happen, for example, if wind is dispersing the population, but the wind speed is not everywhere the same.) There are no additional growth or decay, diffusion, chemotaxis, etc., just pure transport. Circle which of the following PDE's describes the evolution of c:

$$\frac{\partial c}{\partial t} = -x \frac{\partial c}{\partial x} \qquad \frac{\partial c}{\partial t} = -5c \qquad \frac{\partial c}{\partial t} = -5x \frac{\partial c}{\partial x} \qquad \frac{\partial c}{\partial t} = -5c - x \frac{\partial c}{\partial x} \qquad \frac{\partial c}{\partial t} = -5 - \frac{\partial c}{\partial x}$$
$$\frac{\partial c}{\partial t} = -c - x \frac{\partial c}{\partial x}$$

$$\frac{\partial c}{\partial t} = -5c - 5x \frac{\partial c}{\partial x} \qquad \frac{\partial c}{\partial t} = -5c - 5 \frac{\partial c}{\partial x} \qquad \frac{\partial c}{\partial t} = -5c - 5V' \frac{\partial c}{\partial x} \qquad \text{(none of these)}$$

3. Suppose that a population, with density $c(x, t)$ (in one dimension), is being transported with velocity $v = 1$. There is no additional growth or decay, diffusion, chemotaxis, etc., just pure transport. At time $t = 1$, the density is $c(x, 1) = x^2 - 2x + 1$. (Please note that we didn't specify initial conditions at "$t = 0$".)

   (a) Give a formula for $c(x, t)$.

   (b) The density at position $x = 1$ at time $t = 2$ is (circle the right one):

   0 -27 -17 1 -1 2 -2 3 -3 4 -4 27 (none of these)

4. Suppose that $c(x, t)$ is the density of radioactive particles (in one space dimension) being carried east by a wind blowing at 6 mph. The particles decay with a half-life of 4 hours. Suppose that we know that $c(1, t) = \frac{1}{1+t}$ for all t. Find $c(x, t)$ for all x, t.

5. Suppose that a population, with density $c(x, t)$ (in one dimension), is being transported with velocity $v = 5$. There is no additional growth or decay, diffusion, chemotaxis, etc., just pure transport. The initial density is $c(x, 0) = x^2/(1 + x^2)$.

   (a) Give a formula for $c(x, t)$.

   (b) The density at position $x = 12$ at time $t = 2$ is (circle the right one):

   0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 (none of these)

6. Suppose $c(x, t)$ is the density of bacterial population being carried east by a wind blowing at 4 mph. The bacteria reproduce exponentially, with a doubling time of 5 hours.

   (a) Find the density $c(x, t)$ in each of these cases:

   (1) $c(x, 0) \equiv 1$ (2) $c(x, 0) = 2 + \cos x$ (3) $c(x, 0) = \frac{1}{1+x^2}$ (4) $c(x, 1) = 2 + \cos x$

   (5) $c(0, t) \equiv 1$ (6) $c(0, t) = \sin t$ (7) $c(1, t) = \frac{1}{1+e^t}$.

   (b) Sketch the density $c(x, 10)$ at time $t = 10$.

7. Prove the following analog in dimension 3 of the theorem on solutions of the transport equation (the constant velocity v is now a vector): $c(x, y, z, t) = f(x - v_1 t, y - v_2 t, z - v_3 t)e^{-\lambda t}$. (Hint: use $\alpha(x, y, z, t) = c(x + v_1 t, y + v_2 t, z + v_3 t, t)$.)

Problems PDE2: Chemotaxis

1. Give an example of an equation that would model this situation: the speed of movement is an increasing function of the norm of the gradient, but is bounded by some maximal possible speed.

2. We are given this chemotaxis equation (one space dimension) for the concentration of a microorganism (assuming no additional reactions, transport, etc):

$$\frac{\partial c}{\partial t} = \frac{\partial c}{\partial x} \frac{(2\,x - 6)}{\left(2 + (x-3)^2\right)^2} - 2c \left(\frac{(2\,x-6)^2}{\left(2 + (x-3)^2\right)^3} - \frac{1}{\left(2 + (x-3)^2\right)^2} \right).$$

   (a) What is the potential function? (Give a formula for it; just recognize which term is V'.)

   (b) Where (at $x =$?) is the largest amount of food?

3. Analyze these two cases:

   (a) $V'(x_0) > 0, V''(x_0) < 0$
   (b) $V'(x_0) < 0, V''(x_0) > 0$

   in a manner analogous to what was done in the text.

4. Suppose that a population, with density $c(x,t)$ (in one dimension), is undergoing chemotaxis. There are no additional growth or decay, transport, diffusion, etc., just pure chemotaxis.

   The "potential" function is $V(x) = (x-1)^2$, and the proportionality constant "α" is $\alpha = 1$.

   (a) The density satisfies this equation(circle one):

   $$\frac{\partial c}{\partial t} = -c - (x-1)^2 \frac{\partial c}{\partial x} \qquad \frac{\partial c}{\partial t} = -2c - (x-1)\frac{\partial c}{\partial x} \qquad \frac{\partial c}{\partial t} = -2c - 2(x-1)\frac{\partial c}{\partial x}$$
   $$\frac{\partial c}{\partial t} = -c - 2(x-1)\frac{\partial c}{\partial x}$$

   $$\frac{\partial c}{\partial t} = -2c - 2x\frac{\partial c}{\partial x} \qquad \frac{\partial c}{\partial t} = -c - x\frac{\partial c}{\partial x} \qquad \frac{\partial c}{\partial t} = -c + x\frac{\partial c}{\partial x} \qquad \frac{\partial c}{\partial t} = c - x\frac{\partial c}{\partial x}$$
   (none of these)

   (b) Suppose that, at a given time t_0, $c(t_0, x) = 3x + 2$. Then, what is the rate of change $\frac{\partial c}{\partial t}$ of the population, at $t = t_0$ and $x = 3$? (Answer with a number, like "-10".)

5. An organism's chemotactic signaling pathway will not directly respond to the gradient of the external chemoeffector concentration, but, rather, to the gradient of "sensed" concentration, as represented by the degree of activity of appropriate receptors. In this problem, we study one extremely simplified model of chemotaxis based on this idea. We assume that motion is in one dimension.

   Suppose that each receptor may be free or bound by ligand, and we denote the concentrations of free and bound receptors by r_0 and r respectively. We further assume that the ligand binding process equilibrates much faster than both the reaction time of the chemotatic pathway and the speed at which the organism is moving. Thus, we may view $r_0 = r_0(x)$ and $r = r(x)$ as

functions of the current location (more specifically, of the ligand L's concentration $V(x)$ at the location). We take reversible ligand binding

$$L + R_0 \; \underset{\beta}{\overset{\alpha}{\rightleftharpoons}} \; R \, ,$$
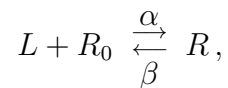
where α, β are two positive constants. Let $r(x) + r_0(x) = 1$ be the (assumed constant) total number of receptors, which we take to be 1 by picking an appropriate unit. By the rapid equilibrium assumption, $V(x)r_0(x) = Kr(x)$, where $K = \beta/\alpha$. Substituting $r_0(x) = 1 - r(x)$, you should solve for $r(x)$ as a function of $V(x)$. Then derive, arguing as before, but using the gradient of r instead of the gradient of V, a flux as follows:

$$J = \alpha \, c \, \frac{K}{(K+V)^2} V'$$

from which we have the equation

$$\frac{\partial}{\partial t}c \; + \; \frac{\partial}{\partial x}\left(\alpha \, c \, \frac{K}{(K+V)^2} V' \right) \; = \; 0$$

for chemotaxis.

Problems PDE3: Diffusion

1. Show that any function of the form

$$c(x, t) = a\, e^{-k^2 t} \sin kx$$

   for some nonzero integer k is a solution of

$$\frac{\partial c}{\partial t} = \frac{\partial^2 c}{\partial x^2}, \quad c(0, t) = c(\pi, t) = 0.$$

2. Under what conditions is there an unbounded (separated form) solution of:

$$\frac{\partial c}{\partial t} = \frac{\partial^2 c}{\partial x^2} + \alpha c, \quad c(0, t) = c(1, t) = 0\ ?$$

   Provide the general form of such solutions. What about boundary condition $\frac{\partial c}{\partial x}(1, t) = 0$?

3. Suppose that $c(x, t)$ denotes the density of a population that is undergoing random motions, with diffusion coefficient $D = 1$ (we ignore reproduction for now). The population lives in a thin tube along the x axis, with endpoints at $x = 0$ and $x = 1$. The endpoint at $x = 0$ is open, and the outside density of bacteria is $c = 5$. The endpoint at $x = 1$ is closed.

   (a) Write down the appropriate diffusion equation, including boundary conditions.

   (b) Find the general form of solutions of the form "constant plus separated": $c(x, t) = 5 + X(x)T(t)$ in which $X(x)T(t)$ is nonzero (i.e., $c(x, t)$ is not constant) and bounded.

4. Suppose that $c(x, t)$ denotes the density of a bacterial population undergoing random motions (diffusion), in dimension 1. The population lives in a thin tube along the x axis, with endpoints at $x = 0$ and $x = \pi/2$. The diffusion constant is $D = 1$. The tube is closed at $x = 0$ and open at $x = \pi/2$, and the outside density of bacteria is $c = 10$.

   (a) Write down the appropriate diffusion equation, including boundary conditions.

   (b) Find the general form of solutions of the form "constant plus separated": $c(x, t) = 10 + X(x)T(t)$ in which $X(x)T(t)$ is nonzero (i.e., $c(x, t)$ is not constant) and bounded.

   (c) Now suppose that we are also told that $c(0, 0) = 12$ and that $\frac{\partial c}{\partial t}(0, 0) = -50$. Find the undetermined constants in the above solution. (Your answer should be explicit, as in "$10 + 3\sin(-3x)e^{2x-7t}$".)

5. Suppose that $c(x, t)$ denotes the density of a bacterial population undergoing random motions (diffusion), in dimension 1. We think of the population as living in a thin tube along the x axis, with endpoints at $x = 0$ and $x = L$, and take the diffusion constant to be $D = 1$ (for simplicity).

   For each of the following descriptions write down the appropriate diffusion equation, including boundary conditions, and then find one solution of the separated form $c(x, t) = X(x)T(t)$ which is bounded and nonzero.

   (a) $x = 0$, $L = 1$, both ends of the tube are open, with the outside density of bacteria being negligible.

   (b) $x = 0$, $L = \pi/2$, end at $x = 0$ open and end at $x = L$ is closed.

(c) $x = 0$, $L = \pi$, both ends are closed.

6. Suppose that $c(x, t)$ denotes the density of a bacterial population undergoing random motions (diffusion), in dimension 1. We suppose that the domain is infinite, $-\infty < x < \infty$, and that, besides diffusion, there is an air flow (in the positive x direction) with constant velocity v. We now let the diffusion coefficient be an arbitrary constant D.

   (a) Explain why we model this by the following equation:

   $$\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial x^2} - v\frac{\partial c}{\partial x}.$$

   (This is called, by the way, an *advection-diffusion* equation, and may be interpreted probabilistically as describing a *random walk with drift*.)

   (b) You will now find the "fundamental solution" of this equation, as follows. Introduce the new variable $z = x - vt$ and the function

   $$\alpha(z, t) = c(z + vt, t).$$

   (You should recognize here the trick which was used in order to solve the transport equation, when there was no diffusion.) Show that α satisfies a diffusion equation, and show therefore how to obtain a solution for c by substituting back into the fundamental solution (Gaussian) for α.

7. In dimension 2, compute the Laplacian in polar coordinates. That is, write

   $$f(r, \varphi, t) = c(r\cos\varphi, r\sin\varphi, t),$$

   so that f is really the same as the function c, but thought of as a function of magnitude, argument, and time. Prove that:

   $$(\nabla^2 c)(r\cos\varphi, r\sin\varphi, t) = \frac{\partial^2 f}{\partial r^2} + \frac{1}{r}\frac{\partial f}{\partial r} + \frac{1}{r^2}\frac{\partial^2 f}{\partial\varphi^2}$$

   (all terms on the RHS evaluated at r, φ, t). Writing f just as c (but remembering that c is now viewed as a function of (r, φ, t)), this means that the diffusion equation in polar coordinates is:

   $$\frac{\partial c}{\partial t} = \frac{\partial^2 c}{\partial r^2} + \frac{1}{r}\frac{\partial c}{\partial r} + \frac{1}{r^2}\frac{\partial^2 c}{\partial\varphi^2}.$$

   Conclude that, for radially symmetric c, the diffusion equation in polar coordinates is:

   $$\frac{\partial c}{\partial t} = \frac{D}{r}\frac{\partial}{\partial r}\left(r\frac{\partial c}{\partial r}\right)$$

   It is also possible to prove that for spherically symmetric c in three dimensions, the Laplacian is $\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial c}{\partial r}\right)$.

8. Show that, under analogous conditions to those in the theorem shown for dimension 1, in dimension d (e.g.: $d = 2, 3$) one has the formula:

   $$\sigma^2(t) = 2dDt + \sigma^2(0)$$

(for $d = 1$, this is the same as previously). The proof will be completely analogous, except that the first step in integration by parts ($uv' = (uv)' - u'v$, which is just the Leibniz rule for derivatives) must be generalized to vectors (use that $\nabla\cdot$ acts like a derivative) and the second step (the Fundamental Theorem of Calculus) should be replaced by an application of Gauss' divergence theorem.

9. We present now an important particular solution of the diffusion equation on $(-\infty, +\infty)$.

   (a) Prove that (for $n = 1$), the following function is a particular solution of the diffusion equation:
   $$c_0(x,t) = \frac{C}{\sqrt{4\pi Dt}} e^{-\frac{x^2}{4Dt}}$$
   (where C is any constant). Also, verify that, indeed for this example, $\sigma^2(t) = 2Dt$.

   (b) In dimension $n = 3$ (or even any other dimension), there is a similar formula. Using a symbolic computation system (e.g., Maple or Mathematica), check that the following function is a solution, for $t > 0$:
   $$c_0(x,t) = \frac{C}{(4\pi Dt)^{3/2}} e^{-\frac{r^2}{4Dt}}$$

   where $r^2 = x_1^2 + x_2^2 + x_3^2$.

   Note that at $t = 0$, these particular solutions are not well-defined, as they tend to a "δ" function. We interpret them as the "spread from a point source". The next problem shows how to use such a solution to generate solutions for arbitrary initial conditions.

10. For any arbitrary continuous function f, show that the function[22]
    $$c(x,t) = \int_{-\infty}^{+\infty} \frac{C}{\sqrt{4\pi Dt}} e^{-\frac{(x-\xi)^2}{4Dt}} f(\xi)\, d\xi$$

    solves the diffusion equation for $t > 0$, and has the initial condition $c(x,0) = f(x)$.

11. Prove that the radially symmetric solution for the diffusion equation
    $$\frac{\partial c}{\partial t} = \frac{D}{r}\frac{\partial}{\partial r}\left(r\frac{\partial c}{\partial r}\right), \quad c(a,t) = 0, \ c(L,t) = c_0.$$

    on a washer (as discussed in the text) is:
    $$c(r) = c_0 \frac{\ln(r/a)}{\ln(L/a)}.$$

12. Prove that the diffusion solution for the spherical shell is:
    $$c(r) = \frac{Lc_0}{L-a}\left(1 - \frac{a}{r}\right).$$

13. This is a problem about diffusion with population growth.

---

[22]This is the *convolution* $c_0 * f$ of f with the "Green's function" c_0 for the PDE

(a) Show if c satisfies (3.1), then, letting $p(x,t) := e^{-\alpha t} c(x,t)$, it follows that $\frac{\partial p}{\partial t} = D\nabla^2 p$.

(b) Show directly (plug-in) that (3.3) is a solution of (3.1).

(c) Show that the equipopulation contours $c =$ constant have $x \approx \beta t$ for large t, where β is some positive constant. That is to say, prove that, if $c(x,t) = c_0$ (for any fixed c_0 that you pick) then

$$\lim_{t\to\infty} \frac{x}{t} = \beta$$

for some β (which depends on the c_0 that you chose). (Hint: solve $\frac{C}{\sqrt{4\pi Dt}} e^{\alpha t - \frac{x^2}{4Dt}} = c_0$ for x and show that $x = \sqrt{a_1 t^2 + a_2 t + a_3 t \ln t}$ for some constants a_i.

14. Suppose that $c(x,t)$ is the density of a bacterial population undergoing random motions (diffusion), and living in a one-dimensional tube with endpoints at $x=0$ and $x=\pi/2$. The bacteria reproduce with rate $\lambda c = c/4$. The tube is closed at $x=0$ and open at $x=\pi/2$, and the outside density of bacteria is $c = 10$. Taking $D=1$ for simplicity:

(a) Write down the appropriate steady-state equation, including boundary conditions.

(b) Find a (non-negative) solution of this steady state equation, in the form $c(x) = aX(x)$, where X is a trigonometric function.

Problems PDE4: Diffusion and travelling waves

1. In the traveling wave model, use separation of variables to obtain this is a solution:

$$U(\xi) \;=\; \frac{1}{2}\left[1 + \tanh\left(\frac{A}{2\sqrt{2}}\xi\right)\right]$$

2. Derive the formula:

$$B^2(2U - 1) + cB - A^2(U - \alpha) \;=\; 0\,.$$

Problems PDE5: General PDE modeling

1. Consider the following equation for (one-dimensional) pure transport with constant velocity:

$$\frac{\partial c}{\partial t} = -2\frac{\partial c}{\partial x}$$

and suppose that this represents a population of bacteria carried by wind.

Modify the model to include the fact that bacteria reproduce at a rate that is a function "$k(p(x,t))$" where $p(x,t)$ is the density of nutrient available near location x at time t. The nutrient gets depleted at a rate proportional to the growth rate of bacteria. The nutrient is not being transported by the wind.

Model this. You will have to write a set of two partial differential equations. The actual form of the function k is not important; it could be, for instance, of a Michaelis-Menten type. Just leave it as "$k(p)$" in your equations.

2. Suppose that a population of bacteria, with density $c(x,t)$ (in one dimension), evolves according to the following PDE:

$$\frac{\partial c}{\partial t} = -\frac{\partial(cV')}{\partial x}$$

where $V(x) = \frac{x^2}{1+x^2}$.

   (a) Sketch a plot of $V(x)$.

   (b) Is this a transport, chemotaxis, or diffusion equation?

   (c) Describe one short sentence what you think the bacteria are doing. (Examples: "the wind is carrying them Westward", "they move toward a food source at $x = 2$", "they are getting away from an antibiotic placed at $x = 5$", "they are moving at random", etc.)

   You are not being asked to solve any equations. Just testing your understanding of the model. Your answer to part (b) should be just one word, and your answer to part (c) should just be one sentence.

3. (a) What might this equation (in one space dimension) represent?:

$$\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial x^2} + \lambda c \qquad (*)$$

   (Provide a short word description, something in the style of "The population density of bacteria undergoing chemotaxis with potential $V(x) = D$ and also being carried by wind westward at λ miles per hour.")

   (b) Suppose that c solves (*), and introduce the new function $b(x,t) = e^{-\lambda t}c(x,t)$. Show that $\frac{\partial b}{\partial t} = D\frac{\partial^2 b}{\partial x^2}$.

   (c) Use the previous part (and what we already learned) to help you find a nonzero solution of (*).

4. Suppose that $c(x,t)$ denotes the density of a bacterial population. For each of the following descriptions, you must provide a differential equation that provides a model of the situation. You are not asked to *solve* any equations. The *only* point of the exercise is to get you used

to "translate" from word descriptions to equations. The values of constants that are used, for velocities, diffusion coefficients, and so on, are arbitrary and have no physical meaning.

We assume dimension 1. We think of the bacteria as living inside a tube of infinite length (perhaps a very long ventilation system). The density is assumed to be uniform in each cross-section, so we model by $c = c(x, t)$ with $-\infty < x < +\infty$.

(a) Bacteria are transported by an air current blowing "east" (towards $x > 0$) at 5 m/s, they grow exponentially with a doubling time of 1 hour.

(b) Bacteria are transported by an air current blowing "west" at 5 m/s, and they grow exponentially with a doubling time of 1 hour.

(c) Bacteria are transported by an air current blowing east at 5 m/s, and they grow exponentially with a doubling time of 1 hour for small populations, but nutrients are restricted, so the density can never be more than $100\mathrm{m}^{-1}$.

(d) Bacteria are transported by an air current blowing east at 5 m/s, and they grow exponentially with a doubling time of 1 hour, and they also move randomly with a diffusion coefficient of 10^{-3}.

(e) Suppose now that there is a source of food at $x = 0$, and bacteria area attracted to this source. Model the potential as $V(x) = e^{-x^2}$.

5. Suppose that $c(x, t)$ denotes the density of a bacterial population. For each of the following descriptions, you must provide a differential equation that provides a model of the situation. You are not asked to *solve* any equations. The *only* point of the exercise is to get you used to "translate" from word descriptions to equations. The values of constants that are used, for velocities, diffusion coefficients, and so on, are arbitrary and have no physical meaning. Now "x" is a vector (x_1, x_2, x_3) in 3-space (bacteria are moving in space).

(a) Bacteria are transported by an air current blowing parallel to the vector $v = (1, -1, 2)$ at 5 m/s, and they grow exponentially with a doubling time of 1 hour.

(b) Bacteria are transported by an air current blowing parallel to the vector $v = (1, -1, 2)$ at 5 m/s, and they grow exponentially with a doubling time of 1 hour for small populations, but nutrients are restricted, so the density can never be more than 100 (in appropriate units).

(c) Bacteria are transported by an air current blowing parallel to the vector $v = (1, -1, 2)$ at 5 m/s, and they also move randomly with a diffusion coefficient of 10^{-3}.

Chapter 4

Stochastic kinetics

4.1 Introduction

Chemical systems are inherently stochastic, as reactions depend on random (thermal) motion. Deterministic models represent an aggregate behavior of the system. They are accurate in much of classical chemistry, where the numbers of molecules are usually expressed in multiples of Avogadro's number, which is $\approx 6 \times 10^{23}$.[1] In such cases, basically by the law of large numbers, the mean behavior is a good description of the system. The main advantage of deterministic models is that they are comparatively easier to study than probabilistic ones. However, they may be inadequate when the "copy numbers" of species, i.e. the numbers of units (ions, atoms, molecules, individuals) are very small, as is often the case in molecular biology when looking at single cells: copy numbers are small for genes (usually one or a few copies), mRNA's (in the tens), ribosomes and RNA polymerases (up to hundreds) and certain proteins may be at low abundances as well. Analogous situations arise in other areas, such as the modeling of epidemics (where the "species" are individuals in various classes), if populations are small. This motivates the study of stochastic models.

We assume that temperature and volume Ω are constant, and the system is well-mixed.

We consider a chemical reaction network consisting of m reactions which involve the n species

$$S_i, \ i \in \{1, 2, \ldots n\}.$$

The reactions $\mathcal{R}_j$, $j \in \{1, 2, \ldots, m\}$ are specified by combinations of reactants and products:

$$\mathcal{R}_j: \quad \sum_{i=1}^{n} a_{ij} S_i \to \sum_{i=1}^{n} b_{ij} S_i \tag{4.1}$$

where the a_{ij} and b_{ij} are non-negative integers, the *stoichiometry coefficients*[2], and the sums are understood informally, indicating combinations of elements. The integer $\sum_{i=1}^{n} a_{ij}$ is the *order* of the reaction $\mathcal{R}_j$. One allows the possibility or zero order, that is, for some reactions j, $a_{ij} = 0$ for all i. This is the case when there is "birth" of species out of the blue, or more precisely, a species is created by what biologists call a "constitutive" process, such as the production of an mRNA molecule by a

---

[1]There is this number of atoms in 12g of carbon-12. A "mole" is defined as the amount of substance of a system that contains an Avogadro number of units.

[2]In Greek, *stoikheion* = element, so "measure of elements"

gene that is always active. Zeroth order reactions may also be used to represent inflows to a system from its environment. Similarly, also allowed is the possibility that, for some reactions j, $b_{ij} = 0$ for all i. This is the case for reactions that involve degradation, dilution, decay, or outflows.

The data in (4.1) serves to specify the stoichiometry of the network. The $n \times m$ *stoichiometry matrix* $\Gamma = \{\gamma_{ij}\}$ has entries:

$$\gamma_{ij} = b_{ij} - a_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m. \tag{4.2}$$

Thus, γ_{ij} counts the net change in the number of units of species S_i each time that reaction $\mathcal{R}_j$ takes place.

We will denote by γ_j the jth column of Γ:

$$\gamma_j = b_j - a_j$$

where[3]

$$a_j = (a_{1j}, \ldots, a_{nj})' \quad \text{and} \quad b_j = (b_{1j}, \ldots, b_{nj})'$$

and assume that no $\gamma_j = 0$ (that is, every reaction changes at least one species).

Stoichiometry information is not sufficient, by itself, to completely characterize the behavior of the network: one must also specify the *rates* at which the various reactions take place. This can be done by specifying "propensity" or "intensity" functions.

We will consider deterministic as well as stochastic models, and propensities take different forms in each case. To help readability, we will use the symbol ρ^σ, possibly subscripted, to indicate stochastic propensities, and $\rho^\#$ and ρ^c to indicate deterministic propensities (for numbers of elements or for concentrations, respectively).

---

[3] prime indicates transpose

4.2 Stochastic models of chemical reactions

Stochastic models of chemical reaction networks are described by a column-vector Markov stochastic process $X = (X_1, \ldots, X_n)'$ which is indexed by time $t \geq 0$ and takes values in $\mathbb{Z}_{\geq 0}^n$. Thus, $X(t)$ is a $\mathbb{Z}_{\geq 0}^n$-valued random variable, for each $t \geq 0$. Abusing notation, we also write $X(t)$ to represent an outcome of this random variable on a realization of the process. The interpretation is:

$$X_i(t) \;=\; \text{number of units of species } i \text{ at time } t \,.$$

One is interested in computing the probability that, at time t, there are k_1 units of species 1, k_2 units of species 2, k_3 units of species 3, and so forth:

$$p_k(t) = \mathbb{P}\left[X(t) = k\right]$$

for each $k \in \mathbb{Z}_{\geq 0}^n$. We call the vector k the *state* of the process at time t.

Arranging the collection of all the $p_k(t)$'s into an infinite-dimensional vector, after an arbitrary order has been imposed on the integer lattice $\mathbb{Z}_{\geq 0}^n$, we have that $p(t) = (p_k)_{k \in \mathbb{Z}_{\geq 0}^n}$ is the discrete probability density (also called the "probability mass function") of $X(t)$.

Biological systems are often studied at "steady state", that is to say after processes have had time to equilibrate. In that context, it is of interest to study the *stationary* (or "equilibrium") density π obtained as the limit as $t \to \infty$ (provided that the limit exists) of $p(t)$. Its entries are the steady state probabilities of being in the state k:

$$\pi_k = \lim_{t \to \infty} p_k(t)$$

for each $k \in \mathbb{Z}_{\geq 0}^n$.

All these probabilities will, in general, depend upon the initial distribution of species, that is, on the $p_k(0)$, $k \in \mathbb{Z}_{\geq 0}^n$, but under appropriate conditions studied in probability theory (ergodicity), the steady state density π will be independent of the initial density.

Also interesting, and often easier to compute, are statistical objects such as the expectation or mean (i.e, the average over all possible random outcomes) of the numbers of units of species at time t:

$$\mathbb{E}\left[X(t)\right] = \sum_{k \in \mathbb{Z}_{\geq 0}^n} p_k(t) k$$

which is a column vector whose entries are the means

$$\mathbb{E}\left[X_i(t)\right] \;=\; \sum_{k \in \mathbb{Z}_{\geq 0}^n} p_k(t) k_i \;=\; \sum_{\ell=0}^{\infty} \ell \sum_{\{k \in \mathbb{Z}_{\geq 0}^n, k_i = \ell\}} p_k(t) \;=\; \sum_{\ell=0}^{\infty} \ell\, p_\ell^{(i)}(t)$$

of the $X_i(t)$'s, where the vector $(p_0^{(i)}(t), p_1^{(i)}(t), p_2^{(i)}(t), \ldots)$ is the marginal density of $X_i(t)$. Also of interest, to understand variability, are the matrix of second moments at time t:

$$\mathbb{E}\left[X(t)X(t)'\right]$$

whose (i,j)th entry is $\mathbb{E}\left[X_i(t)X_j(t)\right]$ and the (co)variance matrix at time t:

$$\mathbb{V}\mathrm{ar}\left[X(t)\right] \;=\; \mathbb{E}\left[\left(X(t) - \mathbb{E}\left[X(t)\right]\right)\left(X(t) - \mathbb{E}\left[X(t)\right]\right)'\right] \;=\; \mathbb{E}\left[X(t)X(t)'\right] - \mathbb{E}\left[X(t)\right]\mathbb{E}\left[X(t)\right]'$$

whose (i,j)th entry is the covariance of $X_i(t)$ and $X_j(t)$, $\mathbb{E}\left[X_i(t)X_j(t)\right] - \mathbb{E}\left[X_i(t)\right]\mathbb{E}\left[X_j(t)\right]$.

4.3 The Chemical Master Equation

A *Chemical Master Equation (CME)* (also known in mathematics as a *Kolmogorov forward equation*) is a system of linear differential equations for the p_k's, of the following form. Suppose given m functions

$$\rho_j^\sigma : \mathbb{Z}_{\geq 0}^n \to \mathbb{R}_{\geq 0}, \quad j = 1, \ldots, m, \quad \text{with } \rho_j^\sigma(0) = 0.$$

These are the *propensity functions* for the respective reactions $\mathcal{R}_j$. As we'll discuss later, the intuitive interpretation is that $\rho_j^\sigma(k)dt$ is the probability that reaction $\mathcal{R}_j$ takes place, in a short interval of length dt, provided that the state was k at the begining of the interval. The CME is:

$$\frac{dp_k}{dt} = \sum_{j=1}^m \rho_j^\sigma(k - \gamma_j)\, p_{k-\gamma_j} - \sum_{j=1}^m \rho_j^\sigma(k)\, p_k, \quad k \in \mathbb{Z}_{\geq 0}^n \tag{4.3}$$

where, for notational simplicity, we omitted the time argument "t" from p, and where we make the convention that $\rho_j^\sigma(k - \gamma_j) = 0$ unless $k \geq \gamma_j$ (coordinatewise inequality). There is one equation for each $k \in \mathbb{Z}_{\geq 0}^n$, so this is an infinite system of linked equations. When discussing the CME, we will assume that an initial probability vector $p(0)$ has been specified, and that there is a unique solution of (4.3) defined for all $t \geq 0$.

Exercise. Suppose that $p(t)$ satisfies the CME. Show that if $\sum_{k \in \mathbb{Z}_{\geq 0}^n} p_k(0) = 1$ then $\sum_{k \in \mathbb{Z}_{\geq 0}^n} p_k(t) = 1$ for all $t \geq 0$. (Hint: first, using that $\rho_j^\sigma(k - \gamma_j) = 0$ unless $k \geq \gamma_j$, observe that, for each $j \in \{1, \ldots, m\}$:

$$\sum_{k \in \mathbb{Z}_{\geq 0}^n} \rho_j^\sigma(k - \gamma_j) p_{k-\gamma_j} = \sum_{k \in \mathbb{Z}_{\geq 0}^n} \rho_j^\sigma(k) p_k$$

and use this to conclude that $\sum_{k \in \mathbb{Z}_{\geq 0}^n} p_k(t)$ must be constant. You may use without proof that the derivative of $\sum_{k \in \mathbb{Z}_{\geq 0}^n} p_k(t)$ with respect to time is obtained by term-by-term differentiation.) $\square$

A different CME results for each choice of propensity functions, a choice that is dictated by physical chemistry considerations. Later, we discuss the special case of mass-action kinetics propensities.

Approximating the derivative $\frac{dp_k}{dt}$ by $\frac{1}{h}[p_k(t+h) - p_k(t)]$, (4.3) means that:

$$p_k(t+h) = \sum_{j=1}^m \rho_j^\sigma(k - \gamma_j)\, h\, p_{k-\gamma_j}(t) + \left(1 - \sum_{j=1}^m \rho_j^\sigma(k)\, h\right) p_k(t) + o(h). \tag{4.4}$$

This equation allows an intuitive interpretaion of the CME, as follows:

The probability of being in state k at the end of the interval $[t, t+h]$ is the sum of the probabilities of the following $m + 1$ events:

- *for each possible reaction $\mathcal{R}_j$, the reaction $\mathcal{R}_j$ happened, and the final state is k, and*

- *no reaction happened, and the final state is k.*

We will justify this interpretation after developing some theory. The discussion will also explain why, for small enough h, the probability that more than one reaction occurs in the interval $[t, t+h]$ is $o(h)$.

We will also introduce the n-column vector:

$$f^{\sigma}(k) := \sum_{j=1}^{m} \rho_j^{\sigma}(k)\, \gamma_j = \Gamma R^{\sigma}(k) \qquad k \in \mathbb{Z}_{\geq 0}^n$$

where $R^{\sigma}(k) = (\rho_1^{\sigma}(k), \ldots, \rho_m^{\sigma}(k))'$.

Interpreting $\rho_j^{\sigma}(k)h$ as the probability that reaction $\mathcal{R}_j$ takes place during an interval of length h (if the current state is k), one may then interpret $f^{\sigma}(k)h$ as the expected change of state during such an interval (since γ_j quantifies the size of the jump if the reaction is $\mathcal{R}_j$). Thus, $f^{\sigma}(k)$ may be thought of as the rate of change of the state, if the state is k.

When studying steady-state properties, we will not analyze convergence of the random variables $X(t)$ as $t \to \infty$. We will simply *define* a (not necessarily unique) steady state distribution $\pi = (\pi_k)$ of the process as any solution of the equations

$$\boxed{\sum_{j=1}^{m} \rho_j^{\sigma}(k - \gamma_j)\, \pi_{k-\gamma_j} - \sum_{j=1}^{m} \rho_j^{\sigma}(k)\, \pi_k = 0, \quad k \in \mathbb{Z}_{\geq 0}^n}.$$

4.3.1 Propensity functions for mass-action kinetics

We first introduce some additional notations. For each $j \in \{1, \ldots, m\}$,

$$A_j = \sum_{i=1}^{n} a_{ij}$$

is the total number of units of all species participating in one reaction of type $\mathcal{R}_j$, the order of $\mathcal{R}_j$.

For each $k = (k_1, \ldots, k_n)' \in \mathbb{Z}_{\geq 0}^n$, we let (recall that a_j denotes the vector $(a_{1j}, \ldots, a_{nj})'$):

$$\binom{k}{a_j} = \prod_{i=1}^{n} \binom{k_i}{a_{ij}}$$

where $\binom{k_i}{a_{ij}}$ is the usual combinatorial number $k_i!/(k_i - a_{ij})!a_{ij}!$, which we define to be zero if $k_i < a_{ij}$.

The most commonly used propensity functions, and the ones best-justified from elementary physical principles, are *ideal mass action kinetics* propensities, defined as follows:

$$\rho_{j,\Omega}^{\sigma}(k) = \frac{c_j}{\Omega^{A_j - 1}} \binom{k}{a_j}, \quad j = 1, \ldots, m. \tag{4.5}$$

The subscript Ω is used for emphasis, even though Ω is a constant, when we want to emphasize how the different rates depend on the volume, but it is omitted when there is no particular interest in the dependence on Ω. The m non-negative constants $c_1, \ldots, c_m$ are arbitrary, and they represent quantities related to the shapes of the reactants, chemical and physical information, and temperature.

4.3.2 Some examples

We will illustrate our subsequent discussions with a few simple but extremely important examples.

mRNA production and degradation

Consider the chemical reaction network consisting of the two reactions $0 \to M$ (formation) and $M \to 0$ (degradation), also represented as:

$$0 \xrightarrow{\alpha} M \xrightarrow{\beta} 0 \tag{4.6}$$

where α and β are the respective rates and mass-action kinetics is assumed. The symbol "0" is used to indicate an empty sum of species.

The application we have in mind is that in which M indicates number of mRNA molecules, and the formation process is transcription from a gene G which is assumed to be at a constant level of activity. (Observe that one could alternatively model the transcription process by means of a reaction "$G \to G + M$" instead of "$0 \to M$", where G would indicate the activity level of the gene G. Since G is neither "created" nor "destroyed" in the reactions, including it in the model is redundant. Of course, if we wanted also to include in our model temporal changes in the activation of G, then a more complicated model would be called for.)

The stoichiometry matrix and propensities are:[4]

$$G = \begin{pmatrix} 1 & -1 \end{pmatrix}, \quad \rho_1^\sigma(k) = \alpha, \quad \rho_2^\sigma(k) = \beta k \tag{4.7}$$

so that

$$f^\sigma(k) = \alpha - \beta k. \tag{4.8}$$

The CME becomes:

$$\frac{dp_k}{dt} = \alpha p_{k-1} + (k+1)\beta p_{k+1} - \alpha p_k - k\beta p_k \tag{4.9}$$

where, recall, the convention is that a term is zero if the subscript is negative. Observe that here $k \in K = \mathbb{Z}_{\geq 0}$ is just a non-negative integer.

We later discuss how to solve the CME for this example. For now, we limit ourselves to a discussion of its steady-state solution.

In general, let π be the steady-state probability distribution obtained by setting $\frac{dp}{dt} = 0$. Under appropriate technical conditions, not discussed here, there is a unique such distribution, and it holds that $\pi_k = \lim_{t \to \infty} p_k(t)$ for each $k \in \mathbb{Z}_{\geq 0}^n$ and every solution $p(t)$ of the CME for an initial condition that is a probability density ($\sum_k p_k(0) = 1$). We may interpret π as the probability distribution of a random variable $X(\infty)$ obtained as the limit of $X(t)$ as $t \to \infty$.

In this example, by definition the numbers π_k satisfy:

$$\alpha \pi_{k-1} + (k+1)\beta \pi_{k+1} - \alpha \pi_k - k\beta \pi_k = 0, \quad k = 0, 1, 2, \ldots \tag{4.10}$$

(the first term is not there if $k = 0$). It is easy to solve recursively for π_k, $k \geq 1$ in terms of π_0, and then use the condition $\sum_k \pi_k(0) = 1$ to find π_0; there results that

$$\pi_k = e^{-\lambda} \frac{\lambda^k}{k!} \tag{4.11}$$

---

[4]Volume dependence is assumed to be already incorporated into α, in this and other examples.

where $\lambda = \frac{\alpha}{\beta}$. In other words, the steady state probability distribution is Poisson distributed with parameter λ.

Exercise. Show, using induction on k, that indeed (4.11) solves (4.10).

Bursts of mRNA production

In an often-studied variation of the above model, mRNA is produced in "bursts" of $r > 1$ (assumed to be a fixed integer) transcripts at a time. This leads to the reactions

$$0 \xrightarrow{\alpha} r\,M\,, \quad M \xrightarrow{\beta} 0 \tag{4.12}$$

with stoichiometry matrix and propensities:

$$\Gamma = (r \quad -1)\,, \quad \rho_1^\sigma(k) = \alpha\,, \quad \rho_2^\sigma(k) = \beta k \tag{4.13}$$

so that

$$f^\sigma(k) = r\alpha - \beta k\,. \tag{4.14}$$

The form of f^σ is exactly the same as in the non-bursting case: the only difference is that the rate α has to be redefined as $r\alpha$. This will mean that the deterministic chemical equation representation is the same as before (up to this redefinition), and, as we will see, the mean of the stochastic process will also be the same (up to redefinition of α). Interestingly, however, we will see that the "noisiness" of the system can be lowered by a factor of up to $1/2$.

Exercise. Write the CME for the bursting model.

A simple dimerization example

Here is another simple example. Suppose that a molecule of A can be produced at constant rate α and degrades when dimerized:

$$0 \xrightarrow{\alpha} A\,, \quad A + A \xrightarrow{\beta} 0 \tag{4.15}$$

which leads to

$$\Gamma = (1 \quad -2)\,, \quad \rho_1^\sigma(k) = \alpha\,, \quad \rho_2^\sigma(k) = \frac{\beta k(k-1)}{2} \tag{4.16}$$

and

$$f^\sigma(k) = \alpha - \beta k(k-1) = \alpha + \beta k - \beta k^2\,. \tag{4.17}$$

Exercise. Write the CME for the dimerization model.

A model of transcription and translation

One of the most-studied models of gene expression is as follows. We consider the reactions for mRNA production and degradation (4.6):

$$0 \xrightarrow{\alpha} M \xrightarrow{\beta} 0$$

together with:

$$M \xrightarrow{\theta} M + P\,, \quad P \xrightarrow{\delta} 0 \tag{4.18}$$
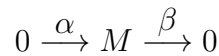
where P represents the protein translated from M. Now

$$\Gamma = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}\,, \quad \rho_1^\sigma(k) = \alpha\,, \ \rho_2^\sigma(k) = \beta k_1\,, \ \rho_3^\sigma(k) = \theta k_1\,, \ \rho_4^\sigma(k) = \delta k_2\,. \tag{4.19}$$

where $k = (k_1, k_2)$ is a vector that counts mRNA and protein numbers respectively, and (writing "(M, P)" instead of $k = (k_1, k_2)$):

$$f^\sigma(M, P) = \begin{pmatrix} \alpha - \beta M \\ \theta M - \delta P \end{pmatrix}. \tag{4.20}$$

Observe that P does not affect M, so the behavior of M will be the same as in the transcription model, and in particular the steady-state distribution of M is Poisson. However, P depends on M, making the problem much more interesting.

Exercise. Write the CME for the transcription/translation model. (Remember that now "k" is a vector (k_1, k_2).)

Remark on FACS: Experimentally estimating the probability distribution of protein numbers

Suppose that we wish to know at what rate a certain gene X is being transcribed under a particular set of conditions in which the cell finds itself. Fluorescent proteins may be used for that purpose. For instance, *green fluorescent protein (GFP)* is a protein with the property that it fluoresces in green when exposed to UV light. It is produced by the jellyfish *Aequoria victoria*, and its gene has been isolated so that it can be used as a *reporter gene*. The GFP gene is inserted (cloned) into the chromosome, adjacent to or very close to the location of gene X, so both are controlled by the same promoter region. Thus, gene X and GFP are transcribed simultaneously and then translated. and so by measuring the intensity of the GFP light emitted one can estimate how much of X is being expressed.

Fluorescent protein methods are particularly useful when combined with *flow cytometry*.[5] Flow Cytometry devices can be used to sort individual cells into different groups, on the basis of characteristics such as cell size, shape, or amount of measured fluorescence, and at rates of up to thousands of cells per second. In this manner, it is possible, for instance, to classify the strength of gene expression in individual cells in a population, perhaps under different sets of conditions.

fluorescent protein construct

cell count versus intensity

---
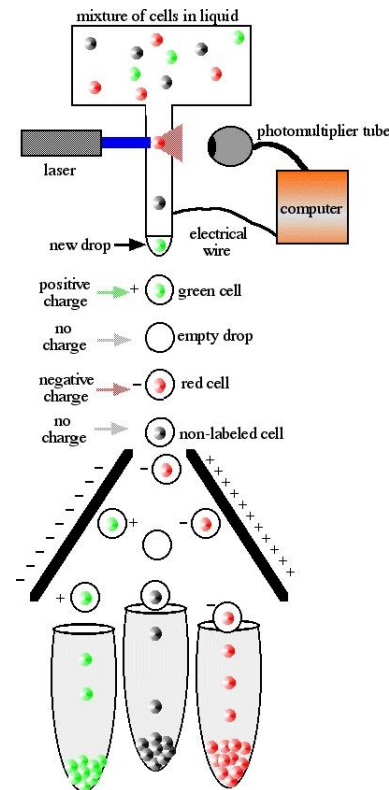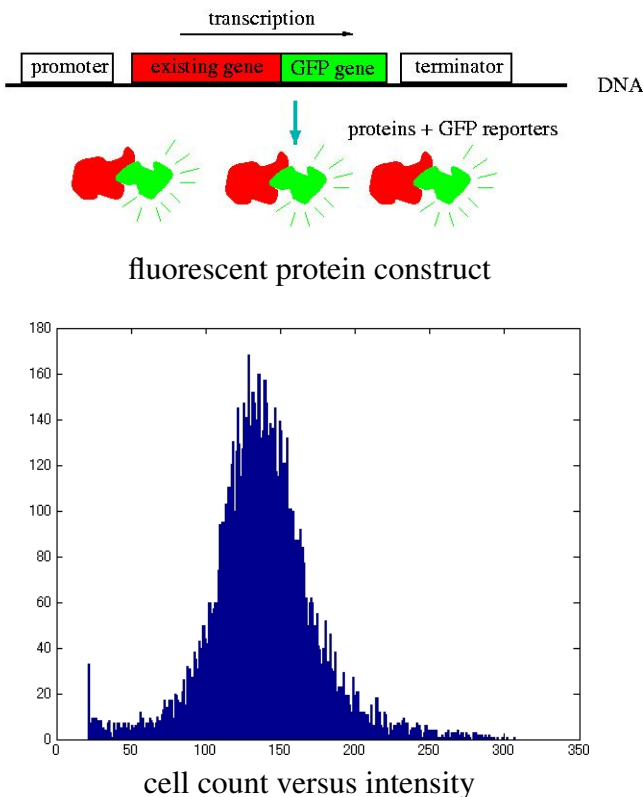
[5]FACS = "fluorescence-activated cell sorting".

4.4 Theoretical background, algorithms, and discussion

The abstract theoretical mathematical background for the CME is as follows.

4.4.1 Markov Processes

Suppose that $\{X(t)\}, t \in [0, \infty)$ is a stochastic process, that is to say a collection of jointly distributed random variables, each of which takes values in a fixed countable set K ($K = \mathbb{Z}_{\geq 0}^n$ in our case).[6]

From now on, we assume that the process is a *continuous-time stationary Markov chain*, meaning that it satisfies the following properties:[7]

- **[Markov]** For any two non-negative real numbers t, h, any function $x : [0, s] \to K$, and any $k \in K$,

$$\mathbb{P}\left[X(t + h) = k \mid X(s) = x(s), 0 \leq s \leq t\right] = \mathbb{P}\left[X(t + h) = k \mid X(t) = x(t)\right].$$

  This property means that $X(t)$ contains all the information necessary in order to estimate the future values $X(T), T \geq t$: additional values from the past do not help to get a better prediction.

- **[Stationarity]** The conditional or *transition* probabilities $\mathbb{P}\left[X(s) = \ell \mid X(t) = k\right]$ depend only on the difference $t - s$. This property, also called homogeneity, means that the probabilities do not change over time.

- **[Differentiability]** With $p_{\ell k}(h) := \mathbb{P}\left[X(t + h) = \ell \mid X(t) = k\right]$ and $p_k(t) := \mathbb{P}\left[X(t) = k\right]$ for every $\ell, k \in K$ and all $t, h \geq 0$, the functions $p_{\ell k}(h)$ and $p_k(t)$ are differentiable in h, t.

Note the following obvious facts:

- $\sum_{\ell \in K} p_{\ell k}(h) = 1$ for every $k \in K$ and $h \geq 0$.

- $p_{\ell k}(0) = \begin{cases} 0 & \text{if } \ell \neq k \\ 1 & \text{if } \ell = k. \end{cases}$

Take any $t, h > 0$, and any $\ell \in K$. Then

$$
\begin{aligned}
p_\ell(t + h) &= \mathbb{P}\left[X(t + h) = \ell\right] = \sum_{k \in K} \mathbb{P}\left[X(t + h) = \ell \ \& \ X(t) = k\right] \\
&= \sum_{k \in K} \mathbb{P}\left[X(t + h) = \ell \mid X(t) = k\right] \times \mathbb{P}\left[X(t) = k\right]
\end{aligned}
$$

---

[6]The more precise notation would be "$X_t(\omega)$", where ω is an element of the outcome space, but we adopt the standard convention of not showing ω. We also do not specify the sample space nor the sigma-algebra of measurable sets which constitute events to which a probability is assigned. If one imposes the requirement that, with probability one, sample paths are continuous from the right and have well-defined limits from the left, a suitable sample space can then be taken to be a space of piecewise constant mappings from $\mathbb{R}_{\geq 0}$ to K.

[7]A subtle fact, usually not mentioned in textbooks, is that conditional probabilities are not always well-defined: "$\mathbb{P}\left[A|B\right] = \mathbb{P}\left[A\&B\right]/\mathbb{P}\left[B\right]$" makes no sense if $\mathbb{P}\left[B\right] = 0$. However, for purposes of our discussions, one may define $\mathbb{P}\left[A|B\right]$ arbitrarily in that case, and no arguments will be affected.

because the events $\{X(t) = k\}$ are mutually exclusive for different k. In other words:

$$p_\ell(t + h) \;=\; \sum_{k \in K} p_{\ell k}(h) p_k(t)\,. \tag{4.21}$$

Similarly. we have[8] the *Chapman-Kolmogorov* equation for the process:

$$p_{\ell\tilde{\ell}}(t + h) \;=\; \sum_{k \in K} p_{\ell k}(t)\, p_{k\tilde{\ell}}(h)\,. \tag{4.22}$$

4.4.2 The jump time process: how long do we wait until the next reaction?

Suppose that $X(t_0) = k$, and consider a time interval $I = [t_0, t_0 + h]$. If $X(t) \neq k$ for some $t \in I$, one says that a "change of state" or an "event" has occurred during the interval, or, for chemical networks, that "a reaction has occurred".

For each $k \in K$ and $h \geq 0$, let:

$$
\begin{aligned}
C_k(h) \;&:=\; \mathbb{P}\left[\text{no reaction occurred on } [t_0, t_0 + h] \,\middle|\, X(t_0) = k\right] \\
&=\; \mathbb{P}\left[X(t) = k \;\forall t \in [t_0, t_0 + h] \,\middle|\, X(t_0) = k\right]
\end{aligned}
$$

(the definition is independent of the particular t_0, by homogeneity). The function $C_k(h)$ is non-increasing on h, and $C_k(0) = 1$. Consider any two $h_1 \geq 0$ and $h_2 \geq 0$. We claim that

$$C_k(h_1 + h_2) \;=\; C_k(h_1) C_k(h_2)\,.$$

Indeed, using the shorthand notation "$X(a, b) = k$" to mean that "$X(t) = k$ for all $t \in [a, b]$", we have:

$$
\begin{aligned}
&\mathbb{P}\left[X(t_0, t_0 + h_1 + h_2) = k \,\middle|\, X(t_0) = k\right] \\
&=\; \mathbb{P}\left[X(t_0, t_0 + h_1 + h_2) = k \;\&\; X(t_0) = k\right] / \mathbb{P}\left[X(t_0) = k\right] \\
&=\; \mathbb{P}\left[X(t_0, t_0 + h_1 + h_2) = k\right] / \mathbb{P}\left[X(t_0) = k\right] \\
&=\; \mathbb{P}\left[X(t_0, t_0 + h_1) = k \;\&\; X(t_0 + h_1, t_0 + h_1 + h_2) = k\right] / \mathbb{P}\left[X(t_0) = k\right] \\
&=\; \mathbb{P}\left[X(t_0, t_0 + h_1) = k\right] \times \mathbb{P}\left[X(t_0 + h_1, t_0 + h_1 + h_2) = k \,\middle|\, X(t_0, t_0 + h_1) = k\right] / \mathbb{P}\left[X(t_0) = k\right] \\
&=\; \mathbb{P}\left[X(t_0, t_0 + h_1) = k\right] \times \mathbb{P}\left[X(t_0 + h_1, t_0 + h_1 + h_2) = k \,\middle|\, X(t_0 + h_1) = k\right] / \mathbb{P}\left[X(t_0) = k\right] \\
&=\; \left(\mathbb{P}\left[X(t_0, t_0 + h_1) = k\right] / \mathbb{P}\left[X(t_0) = k\right]\right) \times \mathbb{P}\left[X(t_0 + h_1, t_0 + h_1 + h_2) = k \,\middle|\, X(t_0 + h_1) = k\right] \\
&=\; C_k(h_1)\, C_k(h_2)
\end{aligned}
$$

(we used the formula $\mathbb{P}[A\&B] = \mathbb{P}[A] \times \mathbb{P}[B|A]$ which comes from the definition of conditional probabilities, as well as the Markov property).

Thus, if we define $c_k(h) = \ln C_k(h)$, we have that $c_k(h_1 + h_2) = c_k(h_1) + c_k(h_2)$, that is, c_k is an additive function. Notice that the functions C_k, and hence also c_k, are monotonic. Therefore each c_k is linear: $c_k(h) = -\lambda_k h$, for some number $\lambda_k \geq 0$.[9] (The negative sign because $c_k(h)$ is the logarithm of a probability, which is a number ≤ 1.) We conclude that $C_k(h) = e^{-\lambda_k h}$.

---

[8]Prove this as an exercise.
[9]Read the Wikipedia article on "Cauchy's functional equation".

In summary:
$$\mathbb{P}\left[\text{no reaction takes place on } [t_0, t_0 + h] \mid X(t_0) = k\right] = e^{-\lambda_k h}$$

from which it follows that

$$\mathbb{P}\left[\text{at least one reaction takes place on } [t_0, t_0 + h] \mid X(t_0) = k\right] = 1 - e^{-\lambda_k h} = \lambda_k h + o(h).$$

A central role in both theory and numerical algorithms is played by the following random variable:

$$\boxed{\mathcal{T}_k := \text{time until the next reaction ("event") will occur, if the current state is } X(t_0) = k}.$$

That is,

if $X(t_0) = k$, an outcome $\mathcal{T}_k = h$ means that the next reaction occurs at time $t_0 + h$.

Observe that, because of the stationary Markov property, $\mathcal{T}_k$ depends only on the current state k, and not on the current time t_0,

If the current time is t_0, then these two events:

- "the next reaction occurs at some time $> h$"

- "no reaction occurs during the interval $[t_0, t_0 + h]$"

are the same. Thus:

$$\mathbb{P}\left[\mathcal{T}_k > h\right] = e^{-\lambda_k h}$$

which means that:

$$\boxed{\textit{the variable } \mathcal{T}_k \textit{ is exponentially distributed with parameter } \lambda_k}.$$

Starting from state k, the time to wait until the Nth subsequent reaction takes place is:

$$\mathcal{T}_{k^{(1)}} + \mathcal{T}_{k^{(2)}} + \ldots + \mathcal{T}_{k^{(N)}}$$

where $k^{(1)} = k$, $k^{(2)}$ is the state reached after the first reaction, $k^{(3)}$ is the state reached after the second reaction (starting from state $k^{(2)}$), and so forth. Note that the choice of which particular "waiting time" random variable $\mathcal{T}_\ell$ is used at each step depends on the past state sequence.

If two or more reactions happen during an interval $[t_0, t_0 + h]$, then $\mathcal{T}_{k^{(1)}} + \mathcal{T}_{k^{(2)}} + \ldots + \mathcal{T}_{k^{(N)}} \le h$ for some N and some sequence of states, so in particular $\mathcal{T}_k + \mathcal{T}_\ell \le h$ for some ℓ. Observe that

$$\mathbb{P}\left[\mathcal{T}_k + \mathcal{T}_\ell \le h\right] \le \mathbb{P}\left[\mathcal{T}_k \le h \,\&\, \mathcal{T}_\ell \le h\right] = \mathbb{P}\left[\mathcal{T}_k \le h\right] \times \mathbb{P}\left[\mathcal{T}_\ell \le h\right] = (\lambda_k h + o(h))(\lambda_\ell h + o(h)) = o(h)$$

because the variables $\mathcal{T}$ are conditioned on the initial state, and are therefore independent.[10] The probability that ≥ 2 reactions happen is upper bounded by $\sum_\ell \mathbb{P}\left[\mathcal{T}_k + \mathcal{T}_\ell \le h\right]$, where the sum is taken over all those states ℓ that can be reached from k after one reaction. We assume from now on that:

*jumps from any given state k can only take place to one of a **finite** number of possible states ℓ*

(4.23)

---

[10]This step in the argument needs to be made more rigorous: one should specify the joint sample space for the $\mathcal{T}$'s.

(as is the case with chemical networks). Thus this sum is finite, and so we can conclude:

$$\mathbb{P}\left[\geq 2 \text{ reactions happen on the interval } [t_0, t_0 + h] \mid X(t_0) = k\right] = o(h).$$

Note that

$$
\begin{aligned}
1 - e^{-\lambda_k h} &= \mathbb{P}\left[\text{some reaction happens on } [t_0, t_0 + h] \mid X(t_0) = k\right] \\
&= \mathbb{P}\left[\text{exactly one reaction happens on } [t_0, t_0 + h] \mid X(t_0) = k\right] \\
&\quad + \mathbb{P}\left[\geq 2 \text{ reactions happen on } [t_0, t_0 + h] \mid X(t_0) = k\right] \quad \left(= o(h)\right)
\end{aligned}
$$

and thus

$$\mathbb{P}\left[\text{exactly one reaction happens on } [t_0, t_0 + h] \mid X(t_0) = k\right] = 1 - e^{-\lambda_k h} + o(h) = \lambda_k h + o(h).$$

For any two states $k \neq \ell$, and any interval $[t_0, t_0 + h]$, $p_{\ell k}(h) = \mathbb{P}\left[X(t + h) = \ell \mid X(t) = k\right]$ is the sum of

$$\mathbb{P}\left[\text{there is a jump from } k \text{ to } \ell \text{ in the interval } [t_0, t_0 + h]\right]$$

plus

$$\mathbb{P}\left[\text{there is no (direct) jump from } k \text{ to } \ell, \text{ but there is a sequence of jumps that take } k \text{ into } \ell\right]$$

and, as the probability of ≥ 2 jumps is $o(h)$, this last probability is $o(h)$. Thus:

$$p_{\ell k}(h) = \mathbb{P}\left[\text{there is a jump from } k \text{ to } \ell \text{ in the interval } [t_0, t_0 + h]\right] + o(h).$$

Assumption (4.23) then implies that

$$p_{\ell k}(h) = o(h) \quad \text{for all but a finite number of states } \ell.$$

4.4.3 Propensities

A key role in Markov process theory is played by the *infinitesimal transition probabilities* defined as follows:

$$q_{\ell k} := \left.\frac{dp_{\ell k}(h)}{dh}\right|_{h=0}$$

Since $p_{\ell k}(h) = o(h)$ for all but a finite number of states ℓ, it follows that, for each k, there are only a finite number of nonzero $q_{\ell k}$'s. In general, $p_{\ell k}(h) = p_{\ell k}(0) + q_{\ell k} h + o(h)$, so, since $p_{\ell k}(0) = 0$ if $\ell \neq k$ and $= 1$ if $\ell = k$:

$$p_{\ell k}(h) = \begin{cases} h q_{\ell k} + o(h) & \text{if } \ell \neq k \\ 1 + h q_{kk} + o(h) & \text{if } \ell = k. \end{cases}$$

Recall that λ_k is the parameter for the exponentially distributed random variable $\mathcal{T}_k$ that gives the time of the next reaction provided that the present state is k. We claim that:

$$q_{kk} = -\lambda_k \quad \text{for all } k.$$

Indeed, $p_{kk}(h) := \mathbb{P}\left[X(t + h) = k \mid X(t) = k\right]$, and this event is the union of the mutually exclusive events "no reaction happened" (which has probability $e^{-\lambda_k h}$) and "two or more reactions happened,

and the end state is again k". This second event has probability $o(h)$, because the probability that more than one reaction happens (even if the final state is different) is already $o(h)$. Thus: $p_{kk}(h) = e^{-\lambda_k h} + o(h)$, which gives $(dp_{kk}/dt)(0) = -\lambda_k$, as claimed.

Note also that, since $\sum_{\ell \in K} p_{\ell k}(h) = 1$ for all h, taking $d/dh|_{h=0}$ gives:

$$\sum_{\ell \in K} q_{\ell k} = 0 \quad \text{or, equivalently,} \quad q_{kk} = -\sum_{\ell \neq k} q_{\ell k} \tag{4.24}$$

and hence also $\lambda_k = \sum_{\ell \neq k} q_{\ell k}$, for every k.

Recall that the Chapman-Kolmogorov equation (4.22) says that $p_{\ell \tilde{\ell}}(t + h) = \sum_{k \in K} p_{\ell k}(t) \, p_{k \tilde{\ell}}(h)$ for all t, h. By definition, $q_{\ell k} = (dp_{\ell k}/dh)(0)$, so taking the derivative with respect to h and evaluating at $h = 0$, we arrive at the *forward Kolmogorov differential equation*

$$\frac{dp_{\ell \tilde{\ell}}}{dt} = \sum_{k \in K} p_{\ell k} \, q_{k \tilde{\ell}} \tag{4.25}$$

which is an equation relating conditional probabilities through the infinitesimal transitions. Similarly, the corresponding equation on probabilities (4.21) is $p_k(t + h) = \sum_{\ell \in K} p_{k\ell}(h) p_\ell(t)$, which leads under differentiation to:

$$\frac{dp_k}{dt} = \sum_{\ell \in K} q_{k\ell} \, p_\ell \,. \tag{4.26}$$

This differential equation is often also called the forward Kolmogorov equation, *and it is exactly the same as the CME* (4.3) $\frac{dp_k}{dt} = \sum_{j=1}^{m} \rho_j^\sigma(k - \gamma_j) \, p_{k-\gamma_j} - \sum_{j=1}^{m} \rho_j^\sigma(k) \, p_k$, where

the propensities $\rho_j^\sigma(k)$ *are, by definition, the infinitesimal transition probabilities* $q_{\ell k}$.

More precisely, consider the m reactions $\mathcal{R}_j$, which produce the stoichiometry changes $k \mapsto k + \gamma_j$ respectively. We define $\rho_j^\sigma(k) = q_{\ell k}$ for $\ell = k + \gamma_j$, $j = 1, \ldots, m$. So:

$$q_{k\ell} = \begin{cases} \rho_j^\sigma(\ell) & \text{if } \ell = k - \gamma_j \text{ for some } j \in \{1, \ldots, m\} \\ -\sum_{\ell \neq k} q_{\ell k} = -\sum_{j=1}^{m} \rho_j^\sigma(k) & \text{if } \ell = k \quad (\text{recall (4.24)}) \\ 0 & \text{otherwise .} \end{cases}$$

Since $\lambda_k = -q_{kk}$,

$$\lambda_k = \sum_{\ell \neq k} q_{\ell k} = \sum_{j=1}^{m} \rho_j^\sigma(k) \,. \tag{4.27}$$

4.4.4 Interpretation of the Master Equation and propensity functions

Since, by definition of the $q_{\ell k}$'s, $p_{k+\gamma_j,k}(h) = q_{k+\gamma_j,k} h + o(h) = \rho_j^\sigma(k) h + o(h)$ and $p_{kk}(h) = 1 + q_{kk}h + o(h) = 1 - \sum_{j=1}^{m} \rho_j^\sigma(k)h + o(h)$,

$$\mathbb{P}\left[X(t+h) = k + \gamma_j \,|\, X(t) = k\right] = \rho_j^\sigma(k) \, h + o(h) \approx \rho_j^\sigma(k) \, h$$

and

$$\mathbb{P}\left[X(t+h) = k \,|\, X(t) = k\right] = 1 - \sum_{j=1}^{m} \rho_j^\sigma(k) \, h + o(h) \approx 1 - \sum_{j=1}^{m} \rho_j^\sigma(k) \, h \,.$$

Since the probability that more than one reaction occurs on an interval of length h is $o(h)$, the probability that $X(t+h) = k + \gamma_j$ is approximately the same as that of $\mathcal{R}_j$ happening in the interval. This justifies the intepretation of the propensity of the reaction $\mathcal{R}_j$ as:

$$\rho_j^\sigma(k)h \approx \text{ probability that the reaction } \mathcal{R}_j \text{ will take place, during a time interval } [t, t+h]$$
$$\text{of (short) duration } h, \text{ if the state was } k \text{ at time } t.$$

In other words, ρ_j^σ is the rate at which the reaction $\mathcal{R}_j$ "fires". This rate depends, obviously, on how many units of the various reactants are present (k). Furthermore, with this interpretation,

$\rho_j^\sigma(k)\, h\, p_k(t) \approx$

$\mathbb{P}\left[\text{reaction } \mathcal{R}_j \text{ takes place during interval } [t, t+h] \mid \text{state was } k \text{ at time } t\right] \times \mathbb{P}\left[\text{state was } k \text{ at time } t\right]$

$= \mathbb{P}\left[\text{state was } k \text{ at time } t \ \& \text{ reaction } \mathcal{R}_j \text{ takes place during interval } [t, t+h]\right],$

and so

$$\sum_{j=1}^m \rho_j^\sigma(k)\, h\, p_k(t)$$

is the probability that the state at time t is k and *some* reaction takes place during the time interval $[t, t+h]$. (Implicitly assuming that these events are mutually exclusive, i.e. at most one reaction can happen, if the time interval is very short.)

Therefore, the second term in (4.4):

$$\left(1 - \sum_{j=1}^m \rho_j^\sigma(k)h\right) p_k(t) \ = \ p_k(t) - \sum_{j=1}^m \rho_j^\sigma(k)\, h\, p_k(t)$$

$\approx \ \mathbb{P}\left[\{\text{initial state was } k\} \setminus \{\text{initial state was } k \text{ and some reaction happens during interval } [t, t+h]\}\right]$

$= \ \mathbb{P}\left[\text{initial state was } k \text{ and no reaction happens during interval } [t, t+h]\right]$

$= \ \mathbb{P}\left[\text{final state is } k \text{ and no reaction happens during interval } [t, t+h]\right]$

where the last equality is true because the events:

$$\text{no reaction happened and the \textit{initial} state was } k$$

and

$$\text{no reaction happened and the \textit{final} state is } k$$

are the same.

On the other hand, regarding the m first terms in (4.4), note that the event:

$$\text{reaction } \mathcal{R}_j \text{ happened and the final state is } k$$

is the same as the event:

$$\text{reaction } \mathcal{R}_j \text{ happened and the initial state was } k - \gamma_j,$$

and the probability of this last event is $\approx \rho_j^\sigma(k - \gamma_j)\, h\, p_{k-\gamma_j}(t)$

In summary, we are justified in interpreting (4.4) as asserting that the probability of being in state k at the end of the interval $[t, t+h]$ is the sum of the probabilities of the following $m+1$ events:

- for each possible reaction $\mathcal{R}_j$, the reaction $\mathcal{R}_j$ happened, and the final state is k, and

- no reaction happened, and the final state is k.

4.4.5   The embedded jump chain

The exponentially distributed variable $\mathcal{T}$ tells what is the waiting time until the next reaction. In order to understand the behavior of the system as a sequence of jumps, one needs, in addition, a random variable that specifies *which* reaction takes place next (or, more generally for Markov processes, to which state is the next transition), *given that* a transition happens.

For each $\ell \neq k$ and h, let $\alpha_{\ell k}(h)$ be the probability that the state is ℓ at time $t + h$, assuming that the initial state is k and that some reaction has happened. If k is not an absorbing state, that is, if transitions out of k are possible, an elementary calculation with conditional probabilities (using that $X(t + h) = \ell$ implies $X(t + h) \neq k$) shows that:[11]

$$\alpha_{\ell k}(h) \;=\; \mathbb{P}\left[X(t + h) = \ell \mid X(t) = k \ \& \ X(t + h) \neq k\right] \;=\; \frac{\mathbb{P}\left[X(t + h) = \ell \mid X(t) = k\right]}{\mathbb{P}\left[X(t + h) \neq k \mid X(t) = k\right]}.$$

Ideally, one would like to compute this expression, but the transition probabilities are hard to obtain. However,

$$\lim_{h \to 0} \alpha_{\ell k}(h) \;=\; \lim_{h \to 0} \frac{p_{\ell k}(h)}{1 - p_{kk}(h)} \;=\; \lim_{h \to 0} \frac{h q_{\ell k} + o(h)}{1 - (1 + h q_{kk} + o(h))} \;=\; -\frac{q_{\ell k}}{q_{kk}} \;=\; \frac{q_{\ell k}}{\sum_{\widetilde{\ell} \neq k} q_{\widetilde{\ell} k}} \;=:\; d_{\ell}^{(k)}.$$

(If k is an absorbing state, the denominators are zero, but in that case we know that $\alpha_{\ell k}(h) = 0$ for all $\ell \neq k$.)

Although in principle only an approximation, it was proved by J.L. Doob[12] that the discrete probability distribution $d_{\ell}^{(k)}$ (for any fixed k, over all $\ell \neq k$), together with the process $\mathcal{T}_k$, characterize a process with the same probability distribution as the original $X(t)$. By itself, the matrix D with entries $d_{\ell}^{(k)}$ is the transition matrix for the discrete-time *embedded Markov chain* or *jump chain* of the process. This discrete chain provides a complete statistical description of the possible sequences of states visited, except that it ignores the actual times at which jumps occur. It is very helpful in theoretical developments, especially in the classification of states ("recurrent", "transient", etc.) of the continuous process.

4.4.6   The stochastic simulation algorithm (SSA)

To understand the behavior of the process $X(t)$, one could attempt to solve the CME (with a known initial $p(0)$) and compute the probability vector $p(t)$. For most problems, this is a computationally very difficult task, starting with the fact that $p(t)$ is an infinite vector. Thus, it is often useful to *simulate* sample paths of the process. Statistics, such as means and variances, can then be obtained by averaging the results of several such simulations.

The naïve approach to simulation is to discretize time into small intervals, and iterate on intervals, randomly deciding at each instant whether a reaction happens. This is not at all an efficient way to proceed: if the discretization is too fine, no reactions will take place in most intervals, and the iteration step is wasted; if the discretization is too gross, we miss fast behaviors. Luckily, there is a far better way to proceed. The basic method[13] for simulating sample paths of CME's is the *stochastic simulation*

---

[11] The calculation is: $\mathbb{P}\left[A|B\&C\right] = \frac{\mathbb{P}\left[A\&B\&C\right]}{\mathbb{P}\left[B\&C\right]} = \frac{\mathbb{P}\left[A\&B\right]}{\mathbb{P}\left[B\&C\right]} == \frac{\mathbb{P}\left[A|B\right]\mathbb{P}\left[B\right]}{\mathbb{P}\left[C|B\right]\mathbb{P}\left[B\right]} = \frac{\mathbb{P}\left[A|B\right]}{\mathbb{P}\left[C|B\right]}$.

[12] "Markoff chains - Denumerable case," *Transactions of the American Mathematical Society* **58**(1945): 455-473.

[13] There are many variants that are often more efficient to implement, but the basic idea is always the same.

algorithm. Also known as the *kinetic Monte Carlo algorithm*[14], it has been probably known and used for a long time, at least since J.L. Doob's work cited earlier, but in its present form was introduced independently by A.B. Bortz, M.H. Kalos, and J.L. Lebowitz[15] and by D.T. Gillespie[16] (the SSA is often called the "Gillespie algorithm" in the systems biology field).

The method is very simple: if the present state is k, first use the random variable $\mathcal{T}_k$ to compute the next-reaction time, and then pick the particular reaction according to the discrete distribution $d_j^{(k)}$, where we are writing $d_j^{(k)}$ instead of $d_{k+\gamma_j}^{(k)}$, for each $j \in \{1, \ldots, m\}$ (all other $d_\ell^{(k)} = 0$). With the notations for propensities used in the CME, we have, for each $J \in \{1, \ldots, m\}$:

$$d_J^{(k)} = \frac{q_{k+\gamma_J,k}}{\sum_{\widetilde{\ell}=k+\gamma_j} q_{\widetilde{\ell},k}} = \frac{\rho_J^\sigma(k)}{\sum_{j=1}^m \rho_j^\sigma(k)} = \frac{\rho_J^\sigma(k)}{\lambda_k}.$$

Generating samples of the exponential random variable $\mathcal{T}_k$ is easy provided that a uniform (pseudo) random number generator is available, like the "rand" function in MATLAB. In general, if U is a uniformly distributed random variable on $[0,1]$, that is, $\mathbb{P}[U < p] = p$ for $p \in [0,1]$, then $\mathcal{T} = -\frac{\ln U}{\lambda}$ is an exponentially distributed random variable with parameter λ, because:

$$\mathbb{P}[\mathcal{T} > t] = \mathbb{P}\left[-\frac{\ln U}{\lambda} > t\right] = \mathbb{P}[U < e^{-\lambda t}] = e^{-\lambda t}.$$

Here is the pseudo-code for the SSA:

Initialization:

1. `inputs: state k, maximal simulation time` $T_{\max}$

2. `set current simulation time` $t := 0$.

Iteration:

1. `compute` $\rho_j^\sigma(k)$`, for each reaction` $\mathcal{R}_j$`,` $j = 1, \ldots, m$

2. `compute` $\lambda := \sum_{j=1}^m \rho_j^\sigma(k)$

3. `if` $\lambda = 0$`, stop (state is an absorbing state, no further transitions are possible)`

4. `generate two uniform random numbers` r_1`,` r_2 `in` $[0,1]$

5. `compute` $T := -\frac{1}{\lambda} \ln r_1$

6. `if` $t + T > T_{\max}$`, stop`

7. `find the index` J `such that` $\frac{1}{\lambda} \sum_{j=1}^{J-1} \rho_j^\sigma(k) \leq r_2 < \frac{1}{\lambda} \sum_{j=1}^J \rho_j^\sigma(k)$

8. `update` $k := k + \gamma_J$

9. `update` $t := t + T$.

Note that, in step 7, the probability that a particular $j = J$ is picked is the same as the length of the interval $[\frac{1}{\lambda} \sum_{j=1}^{J-1} \rho_j^\sigma(k), \frac{1}{\lambda} \sum_{j=1}^J \rho_j^\sigma(k)]$, which is $\frac{1}{\lambda} \rho_J^\sigma(k) = d_J^{(k)}$.

---

[14]In general, "Monte Carlo" methods are algorithms that rely on repeated random sampling to compute their results.

[15]"New algorithm for Monte-Carlo simulations of Ising spin systems," *J. Comput. Phys.* **17**(1975): 10-18.

[16]"A general method for numerically simulating the stochastic time evolution of coupled chemical reactions," *Journal of Computational Physics* **22**(1976): 403-434.

Of course, one will also want to add code to store the sequence of states k and the jump times T, so as to plot sample paths. Note that, in MATLAB, if v is an array with the numbers $\rho_j^\sigma(k)$, then the command "$J = \text{find}(\text{cumsum}(v) > \text{sum}(r_2 * v))$" provides the index J.

Exercise. (1) Implement the SSA in your favorite programming system (MATLAB, Maple, Mathematica). (2) Take the mRNA/protein model described earlier, pick some parameters, and an initial state; now plot many sample paths, averaging to get means and variances as a function of time, as well as steady state means and variances. (3) Compare the latter with the numbers obtained by using theory as described later. □

Remark: An equivalent way to generate the next reaction in the SSA ("direct method") is as follows (the "first reaction method", also discussed by Gillespie): generate m independent exponential random variables $\mathcal{T}_j$, $j = 1, \ldots, m$, with parameters $\lambda_k^{(j)} = \rho_j^\sigma(k)$ respectively –we think of $\mathcal{T}_j$ as indicating when the reaction j would next take place– and pick the "winner" (smallest $\mathcal{T}_j$) as the time (and index) of the next reaction. The same result obtains, because of the following general mathematical fact:[17] if $\mathcal{T}_1, \ldots, \mathcal{T}_m$ are independent exponentially distributed random variables with rate parameters $\mu_1, \ldots, \mu_m$ respectively, then $\mathcal{T} = \min_j \mathcal{T}_j$ is also exponentially distributed, with parameter $\mu = \sum_j \mu_j$. This fact is simple to prove:

$$\mathbb{P}\left[\mathcal{T} > t\right] = \mathbb{P}\left[\mathcal{T}_1 > t \,\&\, \ldots \,\&\, \mathcal{T}_m > t_m\right] = \prod_j \mathbb{P}\left[\mathcal{T}_j > t\right] = \prod_j e^{-\mu_j t} = e^{-\mu t}.$$

Moreover, it is also true that the index J of the variable which achieves the minimum –i.e., the "next reaction"– is a discrete random variable with is distributed according to the law $\mathbb{P}\left[J = p\right] = \mu_p / (\sum_j \mu_j)$.

From a computational point of view, the first reaction method would appear to be less efficient than the direct method, as m random variables have to be generated at each step (compared to just two for the direct method). However, the first reaction method has one advantage: since any given reaction will typically affect only a small number of species, there is no need to re-compute propensities for those indices j for which $\rho_j^\sigma(k)$ has not changed. This observation, together with the use of an indexed priority queue data structure, and a re-use of previously-generated $\mathcal{T}_j$'s, leads to a more efficient algorithm, the "next reaction method" due to M.A. Gibson and J. Bruck.[18]

4.4.7   Interpretation of mass-action kinetics

We explain now, through an informal discussion, how the formula (4.5): $\rho_{j,\Omega}^\sigma(k) = \frac{c_j}{\Omega^{A_j - 1}} \binom{k}{a_j}$ $(j = 1, \ldots, m)$ is derived.

Suppose that the state of the system at time t is $k = (k_1, \ldots, k_n)'$, and we consider an interval of length $0 < h \ll 1$. What is the probability of a reaction $\mathcal{R}_j$ taking place in the interval $[t, t + h]$?

For this reaction to even have a chance of happening, the first requirement is that some subset $\mathcal{S}$ consisting of
$a_{1,j}$ units of species S_1, $a_{2,j}$ units of species S_2, $a_{3,j}$ units of species S_3, ..., $a_{n,j}$ units of species S_n

---

[17]While formally this provides the same numbers, it is not clear *a priori* why reaction times should be independent!

[18]"Efficient exact stochastic simulation of chemical systems with many species and many channels," *J. Phys. Chem. A* **104**(2000): 1876-1889.

come together in some small volume Ω_0 (Ω_0 depends on the physical chemistry of the problem). For the purpose of this discussion, let us call such an event a "collision" and a set of this form a "reactant set" for reaction $\mathcal{R}_j$.

The system is assumed to be "well-mixed", in the sense that species move randomly and fast, thus giving every possible reactant set an equal chance to have a collision.

The basic assumption of mass-action kinetics is that the probability $\rho_j^\sigma(k)h$ that some collision will happen, during a short interval $[t, t+h]$, is proportional to:

- the length h of the interval;

- the probability that a fixed reactant set has a collision; and

- the number of ways in which a reactant set can be picked, if the state is k.

This model implicitly assumes that, if $\Omega_0 \ll \Omega$ (the total volume), then the chance that more than one collision will happen during a short period is much smaller than the probability of just one collision.

There are $\binom{k}{a_j} = \prod_{i=1}^n \binom{k_i}{a_{ij}}$ possible reactant subsets, if the state is k.

Next, we will argue that the probability of a collision, for any one given reactant set $\mathcal{S}$, is $\left(\frac{\Omega_0}{\Omega}\right)^{r-1}$, where $r = A_j$ is the cardinality of $\mathcal{S}$ (the order of the reaction).

From here, one obtains the formula for $\rho_j^\sigma(k)$. (The constant Ω_0 is absorbed into the proportionality constant c_j, which also includes other biophysical information, such as the probability that a reaction takes place when a collision happens, which in turn depends on the collision energy exceeding a threshold value and on the temperature. The Arrhenius equation gives the dependence of the rate constant on the absolute temperature T as $k = Ae^{-E/RT}$, were E is the "activation energy" and R is the gas constant.)

Suppose that $N = \frac{\Omega}{\Omega_0}$ is an integer. (This is a mild hypothesis, if $\Omega \gg \Omega_0$.) Then, the probability of having a collision, for a given reactant set $\mathcal{S}$, is the probability that r balls all land in the same bucket (an "urn" in probability theory) when assigned uniformly at random to one of N buckets.
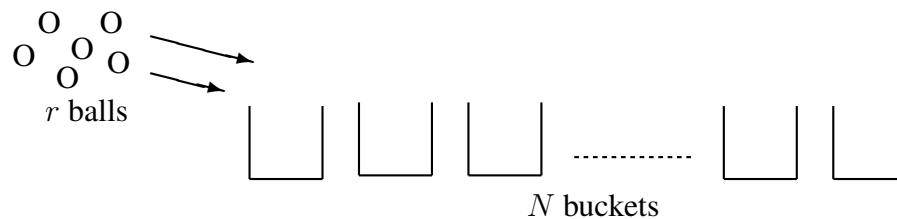
We need to show that this probability is $\left(\frac{1}{N}\right)^{r-1}$. Indeed, the probability that all balls end up in the first bucket is $\left(\frac{1}{N}\right)^r$ (each ball has probability $1/N$ of landing in bucket 1, and the events are independent). The probability that all balls end up in the second bucket is also $\left(\frac{1}{N}\right)^r$, and similarly for all other buckets.

Since the events "all balls land in bucket i" and "all balls land in bucket j" are mutually exclusive for $i \neq j$, the probability of success is $N \times \left(\frac{1}{N}\right)^r = \left(\frac{1}{N}\right)^{r-1}$, which is what we wanted to prove.

The main examples are:

(0) zeroth-order reactions, in which an isolated species is created by means of a process which involve precursors which are not explicitly made a part of the model and which are well-distributed in space; in this case $A_j = 0$ and $\rho_j^\sigma(k)$ is independent of k, so it is just a constant, proportional to the volume;

(1) first-order or monomolecular reactions, in which a single unit of species i is degraded, diluted, decays, flows out, or gets transformed into one or more species; in this case $A_j = 1$ and exactly one a_{ij} is equal to 1 and the rest are zero, so $\rho_j^\sigma(k) = c_j k_i$ (since $\Omega^0 = 1$);

(2) homogeneous second-order (bimolecular) reactions involving two different species S_i and S_ℓ, one unit of each; now there two entries a_{ij} and $a_{\ell j}$ equal to 1, and the rest are zero, $A_j = 2$, and $\rho_j^\sigma(k) = \frac{1}{\Omega} c_j k_i k_\ell$;

(3) homogeneous second-order (bimolecular) reactions involving two units of the same species S_i; now $A_j = 2$ and exactly one a_{ij} is equal to 2 and the rest are zero, so $\rho_j^\sigma(k) = \frac{1}{\Omega} c_j \frac{k_i(k_i-1)}{2}$.

It is frequently argued that at most mono and bimolecular reactions are possible in the real world, since the chance of three or more molecules coming together in a small volume is vanishingly small. In this case, reactions involving multiple species would really consist of a sequence of more elementary bimolecular reactions, involving short-lived, intermediate, species. However, multi-species reactions might still make sense, either as an approximation of a more complicated sequence that occurs very fast, or if molecules are very large compared to the volume, of if the model is one that involves non-chemical substances (for example, in population biology).

4.5 Moment equations and fluctuation-dissipation formula

We next see how to obtain equations for the derivatives of the mean $\mathbb{E}\left[X_i(t)\right]$ and the covariance $\mathbb{V}\text{ar}\left[X(t)\right]$ of $X(t)$, assuming that the probability density of $X(t)$ is given by a CME as in (4.3). No special form needs to be assumed for the propensities, for these theoretical considerations to be valid, but in examples we use mass-action kinetics.

We provide first a very general computation, which we will later specialize to first and second moments. Suppose for this purpose that we have been given a function M which will be, in our two examples, a vector- of matrix-valued function defined on the set of non-negative integers. More abstractly, we take $M : \mathbb{Z}_{\geq 0}^n \to \mathcal{V}$, where $\mathcal{V}$ is any vector space. For first moments (means), we have $\mathcal{V} = \mathbb{R}^n$ and $M(k) = k$. For second moments, $\mathcal{V} = \mathbb{R}^{n \times n}$, the space of all $n \times n$ matrices, and $M(k) = kk'$.[19]

The first goal is to find a useful expression for the time derivative of $\mathbb{E}\left[M(X(t))\right]$. The definition of expectation gives:[20]

$$\mathbb{E}\left[M(X(t))\right] = \sum_{k \in \mathbb{Z}_{\geq 0}^n} p_k(t)\, M(k)$$

because $\mathbb{P}\left[X(t) = k\right] = p_k(t)$. We have:

$$\frac{d}{dt}\mathbb{E}\left[M(X(t))\right] = \sum_{k \in \mathbb{Z}_{\geq 0}^n} \frac{dp_k}{dt}(t)M(k) = \sum_{k \in \mathbb{Z}_{\geq 0}^n}\left(\sum_{j=1}^m \rho_j^\sigma(k-\gamma_j)\, p_{k-\gamma_j} - \sum_{j=1}^m \rho_j^\sigma(k)\, p_k\right)M(k).$$

Note this equality, for each fixed j:

$$\sum_{k \in \mathbb{Z}_{\geq 0}^n} p_{k-\gamma_j}(t)\rho_j^\sigma(k-\gamma_j)M(k) = \sum_{k \in \mathbb{Z}_{\geq 0}^n} p_k(t)\rho_j^\sigma(k)M(k+\gamma_j)$$

(by definition, $\rho_j^\sigma(k-\gamma_j) = 0$ unless $k \geq \gamma_j$, so one may perform a change of variables $\widetilde{k} = k - \gamma_j$). There results:

$$\frac{d}{dt}\mathbb{E}\left[M(X(t))\right] = \sum_{k,j} p_k(t)\rho_j^\sigma(k)M(k+\gamma_j) - \sum_{k,j} p_k(t)\rho_j^\sigma(k)M(k)$$

$$= \sum_{k \in \mathbb{Z}_{\geq 0}^n} p_k(t) \sum_{j=1}^m \rho_j^\sigma(k)\left[M(k+\gamma_j) - M(k)\right].$$

Let us define, for any $\gamma \in \mathbb{Z}_{\geq 0}^n$, the new function $\Delta_\gamma M$ given by $(\Delta_\gamma M)(k) := M(k+\gamma) - M(k)$. With these notations,

$$\frac{d}{dt}\mathbb{E}\left[M(X(t))\right] = \mathbb{E}\left[\sum_{j=1}^m \rho_j^\sigma(X(t))\, \Delta_{\gamma_j} M(X(t))\right]. \tag{4.28}$$

Note that this is *not* an ordinary differential equation for $\mathbb{E}\left[M(X(t)\right]$, because the right-hand side is not, generally, a function of $\mathbb{E}\left[M(X(t))\right]$. In some cases, however, various approximations result in differential equations, as discussed below.

---

[19] As usual, prime indicates transpose, so this is the product of a column vector by a row vector, which is a rank 1 matrix if $k \neq 0$.

[20] Note that this is a deterministic function, not depending on the random outcomes of the process.

Remark. Suppose that M is a polynomial of degree δ_M and that the propensities are polynomials of degree $\leq \delta_\rho$ (the maximal order of reactions, in the mass action case). Then $\Delta_\gamma M$ is a polynomial of degree $\delta_M - 1$, so the monomials appearing inside the expectation have degree $\leq \delta_\rho + \delta_M - 1$. This means that $\frac{d}{dt}\mathbb{E}\left[M(X(t))\right]$ depends on moments of order $\leq \delta_\rho + \delta_M - 1$. Thus, if all reactions have order at most 1, a system of differential equations can obtained for the set of moments of up to any fixed order: the derivative of each moment depends only on equal and lower-order ones, not higher moments. On the other hand, if some reactions have order larger than 1, then $\delta_\rho + \delta_M - 1 > \delta_M$, so in general no clsoed set of equations is available for any finite subset of moments.

4.5.1   Means

For the mean $\mathbb{E}\left[X(t)\right]$, we have $M(k) = k$, so $\Delta_j M(k) = k + \gamma_j - k = \gamma_j$ (a constant function), and thus:

$$\sum_{j=1}^{m} \rho_j^\sigma(X(t))\, \Delta_{\gamma_j} M(X(t)) \;=\; \sum_{j=1}^{m} \rho_j^\sigma(X(t))\, \gamma_j \;=\; f^\sigma(X(t))$$

where, recall, we defined the n-column vector:

$$f^\sigma(k) \;:=\; \sum_{j=1}^{m} \rho_j^\sigma(k)\, \gamma_j \qquad k \in \mathbb{Z}_{\geq 0}^n\,.$$

With these notations, Equation (4.28) specializes to:

$$\boxed{\frac{d}{dt}\mathbb{E}\left[X(t)\right] \;=\; \mathbb{E}\left[f^\sigma(X(t))\right]}\,. \tag{4.29}$$

Recall that $f^\sigma(k)$ can also be written in the form

$$f^\sigma(k) \;=\; \Gamma R^\sigma(k) \tag{4.30}$$

where $R^\sigma(k) = (\rho_1^\sigma(k), \ldots, \rho_m^\sigma(k))'$ and Γ is the stoichiometry matrix.

For mass-action kinetics, the function f^σ is basically the same one[21] that is used in the deterministic differential equation model for the corresponding chemical network. Thus, it is a common mistake to think that the deterministic equation represents an equation that is satisfied by the mean $\mu(t) = \mathbb{E}\left[X(t)\right]$, that is to say, to believe that $d\mu/dt = f^\sigma(\mu)$. However, the precise formula is (4.29). Since expectation of a nonlinear function is generally not the same as the nonlinear function of the expectation[22], (4.29) is, in general, very different from $(d/dt)\mathbb{E}\left[X(t)\right] = f^\sigma(\mathbb{E}\left[X(t)\right])$. One important exception, which permits the replacement $\mathbb{E}\left[f^\sigma(X(t))\right] = f^\sigma(\mathbb{E}\left[X(t)\right])$, is that in which f^σ is an affine function (linear + constant), that is to say if all propensities are affine, which for mass-action kinetics means that all the reactions involve zero or at most one reactant:

$$\boxed{\frac{d}{dt}\mathbb{E}\left[X(t)\right] \;=\; f^\sigma(\mathbb{E}\left[X(t)\right]) \qquad \text{if all reactions are mass-action of order 0 or 1}}\,. \tag{4.31}$$

---

[21]There is just a very minor difference, discussed later, having to do with replacing terms such as "$x(x-1)$" in a second-order homodimerization reaction by the simpler expression x^2.

[22]Example: $\mathbb{E}\left[X^2\right] \neq \mathbb{E}\left[X\right]^2$; in fact, the variance of X is precisely the concept introduced in order to quantify the difference between these two quantities!

On the other hand, even for reactions of arbitrary order, one might expect that Equation (4.31) holds at least approximately provided that the variance of $X(t)$ is small, so that $X(t)$ is almost deterministic. More precisely, one has the following argument.

Let us assume that the function f^σ, which is defined only for non-negative integer vectors, can be extended to a differentiable function, also written as $f^\sigma(x)$, that is defined for all non-negative real numbers x. This is the case with all propensities that are used in practice, such as those arising from mass-action kinetics. Thus, around each vector ξ, we may expand $f^\sigma(x)$ to first-order around $x = \xi$:

$$f^\sigma(x) \; = \; f^\sigma(\xi) \; + \; J(\xi)(x - \xi) \; + \; g_\xi(x - \xi) \tag{4.32}$$

where $J(x)$ is the Jacobian matrix of f^σ evaluated at $x = \xi$ and where g_ξ is a vector function which is $o(|x - \xi|)$. When f is second-order differentiable, the entries g_ξ^i of the vector g_ξ can be expressed as:

$$g_\xi^i(x) \; = \; \frac{1}{2}\,(x - \xi)'\,H_i(\xi)\,(x - \xi) \; + \; o(|x - \xi|^2)$$

where $H_i(\xi)$ is the Hessian of the ith component of the vector field f^σ (the matrix of second order partial derivatives) evaluated at $x = \xi$.

For notational simplicity, let us write μ for means: $\mu(t) = \mathbb{E}\,[X(t)]$. In the particular case that $\xi = \mu(t)$ and $x = X(t)$ (along a sample path), we have:

$$f^\sigma(X(t)) \; = \; f^\sigma(\mu(t)) \; + \; J(\mu(t))\,(X(t) - \mu(t)) \; + \; g_{\mu(t)}(X(t) - \mu(t))\,. \tag{4.33}$$

Now, $J(\mu(t))$ is a deterministic function, so, since the expectation operator is linear,

$$\mathbb{E}\,[J(\mu(t))\,(X(t) - \mu(t))] \; = \; J(\mu(t))\,(\mathbb{E}\,[X(t)] - \mu(t)) \; = \; J(\mu(t))\,(\mathbb{E}\,[X(t)] - \mathbb{E}\,[X(t)]) \; = \; 0\,.$$

Since also $f^\sigma(\mu(t))$ is deterministic, it follows that:

$$\frac{d}{dt}\mathbb{E}\,[X(t)] \; = \; \mathbb{E}\,[f^\sigma(X(t))] \; = \; f^\sigma(\mathbb{E}\,[X(t)]) \; + \; G(t)$$

where

$$G(t) = \mathbb{E}\,\left[g_{\mu(t)}(X(t) - \mu(t))\right]\,. \tag{4.34}$$

This term involves central moments (covariances, etc.) of order ≥ 2.

4.5.2 Variances

For the matrix of second order moments $\mathbb{E}\,[X(t)X(t)']$, we have $M(k) = kk'$, so

$$\Delta_j M(k) \; = \; (k + \gamma_j)(k + \gamma_j)' - kk' \; = \; k\gamma_j' + \gamma_j k' + \gamma_j \gamma_j'$$

and so Equation (4.28) $\frac{d}{dt}\mathbb{E}\,[M(X(t))] = \mathbb{E}\,\left[\sum_{j=1}^m \rho_j^\sigma(X(t))\,\Delta_{\gamma_j} M(X(t))\right]$ specializes to:

$$\frac{d}{dt}\mathbb{E}\,[X(t)X(t)'] \; = \; \mathbb{E}\left[\sum_{j=1}^m \rho_j^\sigma(X(t))\,X(t)\,\gamma_j'\right] + \mathbb{E}\left[\sum_{j=1}^m \rho_j^\sigma(X(t))\,\gamma_j\,X(t)'\right] + \sum_{j=1}^m \mathbb{E}\,\left[\rho_j^\sigma(X(t))\right]\gamma_j\gamma_j'$$

$$\tag{4.35}$$

(note that the $\rho_j^\sigma(X(t))$'s are scalar, and that $X(t)$ and the γ_j's are vectors). Since we had defined $f^\sigma(k) = \sum_{j=1}^m \rho_j^\sigma(k)\,\gamma_j$, the second term in this sum can be written as $\mathbb{E}\left[f^\sigma(X(t))\,X(t)'\right]$. Similarly, the first term is $\mathbb{E}\left[X(t)\,f^\sigma(X(t))'\right]$. The last term can be written in the following useful form.

We introduce the $n \times n$ *diffusion matrix*[23] $B(k) = (B_{pq}(k))$ which has the following entries:

$$B_{pq}(k) \;=\; \sum_{j=1}^m \rho_j^\sigma(k)\,\gamma_{pj}\,\gamma_{qj}\,, \qquad p,q = 1,\ldots,n\,, \tag{4.36}$$

where γ_{pj} is the pth row of the column vector γ_j, that is to say the (p,j)th entry of the stoichiometry matrix Γ, so that $\gamma_{pj}\gamma_{qj}$ is the (p,q)th entry of the matrix $\gamma_j\gamma_j'$. Note that B is an $n \times n$ symmetric matrix. In summary, we can write (4.35) as follows:

$$\boxed{\frac{d}{dt}\mathbb{E}\left[X(t)X(t)'\right] \;=\; \mathbb{E}\left[X(t)\,f^\sigma(X(t))'\right] + \mathbb{E}\left[f^\sigma(X(t))\,X(t)'\right] + \mathbb{E}\left[B(X(t))\right]}\,. \tag{4.37}$$

One interpretation of the entries $\mathbb{E}\left[B_{pq(X(t))}\right]$ is as follows. The product $\gamma_{pj}\gamma_{qj}$ is positive provided both species S_p and S_q change with the same sign (both increase or both decrease) when the reaction $\mathcal{R}_j$ fires. The product is negative if one species increases but the other decreases, when $\mathcal{R}_j$ fires. The absolute value of this product is large if at least one of these two species jumps by a large amount. Finally, the expected value of the coefficient $\rho_j^\sigma(k)$ quantifies the rate at which the corresponding reaction takes place. In this manner, $\mathbb{E}\left[B_{pq(X(t))}\right]$ contributes toward an instantaneous change in the correlation between species S_p and S_q.

An equation for the derivative of the variance is easily obtained from here. By definition, $\mathbb{V}\mathrm{ar}\left[X(t)\right] = \mathbb{E}\left[X(t)X(t)'\right] - \mathbb{E}\left[X(t)\right]\mathbb{E}\left[X(t)\right]'$, so we need to compute the derivative of this last term. For a vector function $v = v(t)$, $(d/dt)vv' = v(dv/dt)' + (dv/dt)v'$, so with $dv/dt = \frac{d}{dt}\mathbb{E}\left[X(t)\right] = \mathbb{E}\left[f^\sigma(X(t))\right]$ from (4.29),

$$\boxed{\frac{d}{dt}\mathbb{V}\mathrm{ar}\left[X(t)\right] \;=\; \mathbb{E}\left[(X(t)-\mu(t))\,f^\sigma(X(t))'\right] + \mathbb{E}\left[f^\sigma(X(t))\,(X(t)-\mu(t))'\right] + \mathbb{E}\left[B(X(t))\right]}$$
$$\tag{4.38}$$

where we wrote $\mu(t) = \mathbb{E}\left[X(t)\right]$ for clarity.

Exercise. Show that an alternative way of writing the third term in the right-hand side of (4.38) is as follows:

$$\Gamma \,\mathrm{diag}\left(\mathbb{E}\left[\rho_1^\sigma(X(t))\right],\ldots,\mathbb{E}\left[\rho_m^\sigma(X(t))\right]\right)\Gamma' \tag{4.39}$$

(where "$\mathrm{diag}\left(r_1,\ldots,r_m\right)$" means a diagonal matrix with entries r_i in the diagonal). $\qquad\square$

The first-order Taylor expansion of f^σ, $f^\sigma(X(t)) = f^\sigma(\mu(t)) + J(\mu(t))(X(t)-\mu(t)) + g_{\mu(t)}(X(t)-\mu(t))$, given in (4.33), can be substituted into the term $\mathbb{E}\left[f^\sigma(X(t))\,(X(t)-\mu(t))'\right]$ in the formula (4.38) for the covariance, giving (dropping the arguments "t" for readability):

$$\mathbb{E}\left[f^\sigma(X)(X-\mu)'\right] \;=\; \mathbb{E}\left[f^\sigma(\mu)(X-\mu)'\right] + \mathbb{E}\left[J(\mu)(X-\mu)(X-\mu)'\right] + \mathbb{E}\left[g_\mu(X-\mu)(X-\mu)'\right]$$
$$=\; J(\mu)\mathbb{V}\mathrm{ar}\left[X\right] + \mathbb{E}\left[g_\mu(X-\mu)(X-\mu)'\right]$$

---

[23]Normally, "diffusion" is interpreted in a spatial sense. Here it is thought of, instead, as diffusion in "concentration space".

where we used that $f^\sigma(\mu(t))$ and $J(\mu(t))$ are deterministic and that $\mathbb{E}[X - \mu] = 0$. Similarly,

$$\mathbb{E}\left[(X - \mu)f^\sigma(X)'\right] = \mathbb{V}\text{ar}[X] J(\mu)' + \mathbb{E}\left[(X - \mu) g_\mu(X - \mu)'\right]$$

(the covariance matrix is symmetric, so there is no need to transpose it). Therefore,

$$\frac{d}{dt}\mathbb{V}\text{ar}[X(t)] = \mathbb{V}\text{ar}[X(t)] J(\mu(t))' + J(\mu(t))\mathbb{V}\text{ar}[X(t)] + \mathbb{E}[B(X(t))] + \alpha(t) \tag{4.40}$$

where $\alpha(t) = \mathbb{E}\left[(X(t) - \mu(t)) g_{\mu(t)}(X(t) - \mu(t))' + (X(t) - \mu(t)) g_{\mu(t)}(X(t) - \mu(t))'\right]$. Dropping the term $\alpha(t)$, one has the *fluctuation-dissipation* formula:

$$\boxed{\frac{d}{dt}\mathbb{V}\text{ar}[X(t)] \approx \mathbb{V}\text{ar}[X(t)] J(\mu(t))' + J(\mu(t))\mathbb{V}\text{ar}[X(t)] + \mathbb{E}[B(X(t))] \quad \text{(FD)}} . \tag{4.41}$$

If the higher-order moments of $X(t)$ are small, one may be justified in making this approximation, because $\alpha(t)$ is $o(|X(t) - \mu(t)|^2)$, while the norm of the covariance matrix is $O(|X(t) - \mu(t)|^2)$.

Equation (4.41) is sometimes called the *mass fluctuation kinetics* equation, and the term "fluctuation-dissipation" is used for a slightly different object, as follows. Suppose that we expand $B(x)$ as a Taylor series around the mean $\mathbb{E}[X(t)]$. Arguing as earlier, we have that $\mathbb{E}[B(X(t))] = B(\mathbb{E}[X(t)]) + o(|X(t) - \mu(t)|)$. This suggests replacing the last term in (FD) by $B(\mathbb{E}[X(t)])$.

4.5.3   Reactions or order ≤ 1 or ≤ 2

The special case in which f^σ is a polynomial of degree two is arguably the most general that often needs to be considered. (Recall the discussion about reactions of order > 2.) In this case, the function g_ξ in (4.32) is a vector field that is quadratic on the coordinates of $X(t) - \mu(t)$, with *constant* coefficients, because the Hessian of a quadratic polynomial is constant. The expectations of such expressions are the covariances $\mathbb{C}\text{ov}[X_i(t), X_j(t)]$ (variances if $i = j$). So, $G(t)$ is a linear function L of the n^2 entries of $\mathbb{V}\text{ar}[X(t)]$. The linear function L can be easily computed from the second derivatives of the components of f^σ. Similarly, as the entries of the diffusion matrix (4.36) are polynomials of degree equal to the largest order of the reactions, when all reactions have order ≤ 2 the term $\mathbb{E}[B(X(t))]$ is an affine linear function of the entries of $\mathbb{E}[X(t)]$ and $\mathbb{V}\text{ar}[X(t)]$, which we write as $H_0 + H_1\mathbb{E}[X(t)] + H_2\mathbb{V}\text{ar}[X(t)]$. Thus:

For mass-action kinetics and all reactions of order at most 2, the fluctuation-dissipation equation says that the mean $\mu(t) = \mathbb{E}[X(t)]$ and covariance matrix $\Sigma(t) = \mathbb{V}\text{ar}[X(t)]$ satisfy

$$d\mu/dt = f^\sigma(\mu) + L\Sigma \tag{4.42a}$$
$$d\Sigma/dt \approx \Sigma J(\mu)' + J(\mu)\Sigma + H_0 + H_1\mu + H_2\Sigma \tag{4.42b}$$

(where the "approximate" sign indicates that α, which involves third-order moments, because $g_{\mu(t)}$ is quadratic, was dropped). Moreover, the function $J(\mu(t))$ is linear in $\mu(t)$.

The FD formula is *exact* for zero- and first-order mass-action reactions, because in that case the Hessian and thus $g_{\mu(t)}$ are zero, so also $\alpha(t) \equiv 0$. Moreover, in this last case the entries $B_{pq}(k) = \sum_{j=1}^m \rho_j^\sigma(k)\gamma_{pj}\gamma_{qj}$ of the diffusion matrix are also affine, so that the last term is just $B(\mathbb{E}[X(t)])$. It is worth emphasizing this fact:

For mass-action kinetics and all reactions of order zero or one, the mean $\mu(t) = \mathbb{E}\left[X(t)\right]$ ***and covariance matrix*** $\Sigma(t) = \mathbb{V}ar\left[X(t)\right]$ ***are solutions of the coupled system of differential equations***

$$\dot{\mu} \;=\; f^{\sigma}(\mu) \tag{4.43a}$$

$$\dot{\Sigma} \;=\; \Sigma\,J' \,+\, J\,\Sigma \,+\, B(\mu) \tag{4.43b}$$

and in this case J does not depend on μ, because J is a constant matrix, being the Jacobian of an affine vector field. Also,

$$B(\mu) = \Gamma\,\text{diag}\left(\rho_1^{\sigma}(\mu), \ldots, \rho_m^{\sigma}(\mu)\right)\,\Gamma' \tag{4.44}$$

in the case of order ≤ 1.

Note that (4.43) is a set of $n + n^2$ linear differential equations. Since covariances are symmetric, however, one can equally well restrict to the equations on the diagonal and upper-triangular part of Σ, so that it is sufficient to solve $n + n(n+1)/2$ equations.

The term "fluctuation-dissipation" is used because the first two terms for Σ may be though of as describing a "dissipation" of initial uncertainty, while the last term can be thought of as a "fluctuation" due to future randomness. To understand the dissipation component, let's discuss what would happen if the fluctuation term were not there. Then (FD) is a linear differential equation on $\mathbb{V}ar\left[X(t)\right]$ (a "Lyapunov equation" in control theory). Given the initial variance $\mathbb{V}ar\left[X(0)\right]$, a solution can be computed. This solution is identically zero when $X(0)$ is perfectly known (that is, $p(0)$ has exactly one nonzero entry), because $\mathbb{V}ar\left[X(0)\right] = 0$ in that case. But even for nonzero $\mathbb{V}ar\left[X(0)\right]$, and under appropriate stability conditions one would have that $\mathbb{V}ar\left[X(t)\right] \to 0$ as $t \to \infty$. If a matrix J has eigenvalues with negative real part, then the operator $P \mapsto PJ' + JP$ on symmetric matrices has all eigenvalues also with negative real part.[24] So if $\mu(t)$ is approximately constant and the linearization of the differential equation for the mean is stable, the equation for the variance will be, too. Since in general the matrices $J(\mu(t))$ depend on t, this argument is not quite correct, but it provides the basic intuition for the term "dissipation".

---

[24]This is because the eigenvalues of this operator are the sums of pairs of eigenvalues of J; see e.g. the author's control theory textbook.

4.6 Generating functions

We next discuss how to use generating functions in order to (1) find solutions of the CME, or at least (2) find differential equations satisfied by moments. Often only simple problems can be solved explicitly with this technique, but it is nonetheless a good source of theoretical insight.

We assume that $p(t)$, an infinite vector function of time indexed by $k \in K = \mathbb{Z}_{\geq 0}^n$, is a solution of the CME (4.3):

$$\frac{dp_k}{dt} = \sum_{j=1}^{m} \rho_j^\sigma(k - \gamma_j)\, p_{k-\gamma_j} - \sum_{j=1}^{m} \rho_j^\sigma(k)\, p_k\,.$$

The *(probability) generating function* $P(z,t)$ is a scalar-valued function of time $t \geq 0$ and of n auxilliary variables $z = (z_1, \ldots, z_n)$ (which may be thought of as complex variables), defined as follows:

$$P(z,t) := \mathbb{E}\left[z^X\right] = \sum_{k \in K} p_k(t)\, z^k \tag{4.45}$$

where we denote $z^k := z_1^{k_1} \ldots z_n^{k_n}$ and $z_i^0 = 1$. As the $p_k(t)$'s are non-negative and add up to one, the series is convergent for $z = 1$ (we write the vector $(1, \ldots, 1)$ as "1" when clear from the context):

$$P(1,t) = 1 \qquad \text{for all } t \geq 0. \tag{4.46}$$

Moments of arbitrary order can be computed once that P is known. For example,

$$\left.\frac{\partial P(z,t)}{\partial z}\right|_{z=1} = \mathbb{E}\left[X(t)\right],$$

where we interpret the above partial derivative as the vector $\left(\left.\frac{\partial P(t,z)}{\partial z_1}\right|_{z=1}, \ldots, \left.\frac{\partial P(t,z)}{\partial z_m}\right|_{z=1}\right)'$. Also,

$$\left.\frac{\partial^2 P(z,t)}{\partial z_i \partial z_j}\right|_{z=1} = \begin{cases} \mathbb{E}\left[X_i(t)X_i(t)\right] & \text{if } i \neq j \\ \mathbb{E}\left[X_i(t)^2\right] - \mathbb{E}\left[X_i(t)\right] & \text{if } i = j\,. \end{cases}$$

Note that $\mathbb{V}\mathrm{ar}\left[X(t)\right]$ can be computed from these formulas.

Exercise. Prove the above two formulas. □

We remark that there are other power series than are often associated to P, especially the *moment generating function*[25]

$$M(\theta,t) := \mathbb{E}\left[e^{\theta X}\right] = \sum_{k \in K} p_k(t)\, e^{\theta k}$$

where we define $e^q = e^{q_1} \ldots e^{q_n}$.

Of course, actually computing $P(z,t)$ from its definition is not particularly interesting, since the whole purpose of using generating functions is to gain information about the unknown $p_k(t)$'s. The idea, instead, is to use the knowledge that $p(t)$ satisfies an infinite system of ordinary differential equations in order to obtain a *finite* set of *partial* differential equations for P. Sometimes these PDE's can be solved, and other times just the form of the PDE will be enough to allow computing ODE's for moments. We illustrate both of these ideas next, through examples.

---

[25]The terminology arises from the fact that the coefficients of the Taylor expansions of P and M, at $z = 0$ and $\theta = 0$, give the probabilities and moments, respectively.

Let us start with the mRNA example given by the reactions in (4.6), $0 \xrightarrow{\alpha} M \xrightarrow{\beta} 0$, for which (cf. (4.7)-(4.9)) $G = (1, -1)$, $\rho_1^\sigma(k) = \alpha$, $\rho_2^\sigma(k) = \beta k$, $f^\sigma(k) = \alpha - \beta k$, and the CME is

$$\frac{dp_k}{dt} = \alpha p_{k-1} + (k+1)\beta p_{k+1} - \alpha p_k - k\beta p_k .$$

Let us compute now a PDE for $P(z, t)$. For simplicity, from now on we will write $\frac{\partial}{\partial t} P$ as P_t and $\frac{\partial}{\partial z} P$ as P_z.

By definition, $P(z, t) = \sum_{k=0}^{\infty} p_k(t) z^k$, so

$$P_t = \sum_{k=0}^{\infty} \frac{dp_k}{dt} z^k = \alpha \sum_{k=1}^{\infty} p_{k-1} z^k + \beta \sum_{k=0}^{\infty} (k+1) p_{k+1} z^k - \alpha \sum_{k=0}^{\infty} p_k z^k - \beta \sum_{k=1}^{\infty} k p_k z^k \quad (4.47)$$

where we started the first sum at 1 because of the convention that $p_{-1} = 0$, and the last at 1 because for zero we have a factor $k = 0$. The third sum in the right-hand side is just P; the rest are:

$$\sum_{k=1}^{\infty} p_{k-1} z^k = z \sum_{k=1}^{\infty} p_{k-1} z^{k-1} = z \sum_{k=0}^{\infty} p_k z^k = zP$$

$$\sum_{k=0}^{\infty} (k+1) p_{k+1} z^k = \sum_{k=1}^{\infty} k p_k z^{k-1} = P_z$$

$$\sum_{k=1}^{\infty} k p_k z^k = z \sum_{k=1}^{\infty} k p_k z^{k-1} = zP_z .$$

Thus, P satisfies:

$$P_t = \alpha z P + \beta P_z - \alpha P - \beta z P_z \quad (4.48)$$

which can also be written as

$$P_t = (z - 1)(\alpha P - \beta P_z) . \quad (4.49)$$

To obtain a unique solution, we need to impose an initial condition, specifying $p(0)$, or equivalently $P(z, 0)$.

(Recall from Equation (4.46) that we also have the boundary condition $P(1, t) = 1$ for all t, because $p(t)$ is a probability distribution.)

Let us say that we are interested in the solution that starts with $M = 0$: $p_0(0) = \mathbb{P}[M(0) = 0] = 1$ and $p_k(0) = \mathbb{P}[M(0) = k] = 0$ for all $k > 0$. This means that $P(z, 0) = \sum_{k=1}^{\infty} p_k(0) z^k = 1$.

Equation (4.49) is a first-order PDE for P. Generally speaking, such PDE's can be solved by the "method of characteristics." Here we simply show that the following guess, which satisfies $P(1, t) = 0$ and $P(z, 0) = 1$:

$$P(z, t) = e^{\frac{\alpha}{\beta}\left(1 - e^{-\beta t}\right)(z-1)} \quad (4.50)$$

is a solution.[26] Indeed, note that, with this definition,

$$P_t(z, t) = \alpha e^{-\beta t}(z - 1) P(t, z) \quad (4.51)$$

---

[26]To be added: solution by characteristics and proof of uniqueness.

$$P_z(z,t) = \frac{\alpha}{\beta}(1 - e^{-\beta t}) \, P(t,z) \tag{4.52}$$

so:

$$P_t = (z-1)\,\alpha e^{-\beta t} P = (z-1)\left[\alpha P - \beta\frac{\alpha}{\beta}(1-e^{-\beta t})P\right] = (z-1)\,(\alpha P - \beta P_z)$$

as claimed.

Once that we have obtained the formula (4.50) for $P(z,t)$, we can expand it in a Taylor series in order to obtain $p_k(t)$. For example, for $k = 1$ we have:

$$\mathbb{P}\left[X(t) = 1\right] = p_1(t) = P_z(0,t) = \frac{\alpha}{\beta}(1 - e^{-\beta t})P(0,t) = \frac{\alpha}{\beta}(1 - e^{-\beta t})e^{-\frac{\alpha}{\beta}\left(1-e^{-\beta t}\right)}.$$

We can also compute moments, for example

$$\mu(t) = \mathbb{E}\left[X(t)\right] = P_z(1,t) = \frac{\alpha}{\beta}\left(1 - e^{-\beta t}\right)P(1,t) = \frac{\alpha}{\beta}\left(1 - e^{-\beta t}\right).$$

As mentioned above, even without solving the PDE for P, one may obtain ODE's for moments from it. For example we have:[27]

$$
\begin{aligned}
\dot{\mu} &= \left.\frac{\partial}{\partial t}\frac{\partial}{\partial z}\right|_{z=1} P = \left.\frac{\partial}{\partial z}\right|_{z=1} P_t = \left.\frac{\partial}{\partial z}\right|_{z=1} (z-1)\,(\alpha P - \beta P_z) \\
&= (\alpha P - \beta P_z) + (z-1)\,(\alpha P_z - \beta P_{zz})\big|_{z=1} \\
&= \alpha - \beta P_z(1,t) = \alpha - \beta\mu.
\end{aligned}
$$

Since every reaction has order 0 or 1, this equation for the mean is the same as the deterministic equation satisfied by concentrations.

Exercise. Use the PDE for P to obtain an ODE for the variance, following a method similar to that used for the mean.

Still for the mRNA example, let us compute the generating function $Q(z)$ of the steady state distribution π obtained by setting $\frac{dp}{dt} = 0$. At steady state, that is setting $P_t = 0$, we have that $(z-1)\,(\alpha Q - \beta Q_z) = 0$, so $\alpha Q - \beta Q_z = 0$, or equivalently $Q_z = \lambda Q$, where $\lambda = \frac{\alpha}{\beta}$. Thus, $Q(z) = ce^{\lambda z}$ for some constant c. Since π is a probability distribution, $Q(1) = 1$, and so $c = e^{-\lambda}$, and thus we conclude:

$$Q(z) = e^{-\lambda}e^{\lambda z} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}z^k.$$

Therefore, since by definition $Q(z) = \sum_{k=0}^{\infty} q_k z^k$, it follows that

$$q_k = e^{-\lambda}\frac{\lambda^k}{k!}$$

and we yet again have recovered the fact that the steady-state distribution is that of a Poisson random variable with parameter λ.

---

[27] Using $\frac{\partial}{\partial t}\frac{\partial}{\partial z} = \frac{\partial}{\partial z}\frac{\partial}{\partial t}$.

4.7 Examples computed using the fluctuation-dissipation formula

Consider again the mRNA example given by the reactions in (4.6), $0 \xrightarrow{\alpha} M \xrightarrow{\beta} 0$, for which (cf. (4.7)-(4.9)) $G = (1, -1)$, $\rho_1^\sigma(k) = \alpha$, $\rho_2^\sigma(k) = \beta k$, $f^\sigma(k) = \alpha - \beta k$. Since the reactions are of order 0 and 1, the FD formula is exact, so that the mean and variance $\mu(t)$ and $\Sigma(t)$ satisfy (4.43): $\dot\mu = f^\sigma(\mu)$, $\dot\Sigma = \Sigma J' + J\Sigma + B(\mu)$. Here both μ and Σ are scalar variables. The Jacobian of f^σ is $J = -\beta$. The diffusion term is

$$B(\mu) = \sum_{j=1}^{2} \rho_j^\sigma(\mu)\gamma_{1j}\gamma_{1j} = \alpha 1^2 + \beta\mu(-1)^2 = \alpha + \beta\mu,$$

so that the FD equations become:

$$\dot\mu = \alpha - \beta\mu \tag{4.53a}$$
$$\dot\Sigma = -2\beta\Sigma + \alpha + \beta\mu. \tag{4.53b}$$

Note that the equation for the mean is the same that we derived previously using the probability generating function. There is a unique steady state for this equation, given by $\mu = \alpha/\beta = \lambda$ (the parameter of the Poisson random variable $X(\infty)$) and, solving $-2\beta\Sigma + \alpha + \beta\mu = 0$:

$$\Sigma = \frac{\alpha + \beta\mu}{2\beta} = \frac{\alpha}{\beta} = \lambda$$

which is, of course, consistent with the property that the variance and mean of a Poisson random variable are the same.

Exercise. Derive the variance equation from the probability generating function, and show that the same result is obtained.

Exercise. Solve explicitly the linear differential equations (4.53). (Use matrix exponentials, or variation of parameters.)

One measure of how "noisy" a scalar random variable X is, is the ratio between its standard deviation $\sigma = \sqrt{\Sigma}$ and its mean, called the *coefficient of variation*:

$$\mathbf{cv}[X] := \frac{\sigma[X]}{\mathbb{E}[X]}$$

(only defined if $\mathbb{E}[X] \neq 0$).

This number may be small even if the variance is large, provided that the mean is large. It represents a "relative noise" and is a "dimensionless" number, thus appropriate, for example, when comparing objects measured in different units.[28]

For a Poisson random variable X with parameter λ, $\mathbb{E}[X] = \lambda$ and $\sigma[X] = \sqrt{\lambda}$, so $\mathbf{cv}[X] = 1/\sqrt{\lambda}$.

Next, we return to the mRNA bursting example given by the reactions in (4.12), $0 \xrightarrow{\alpha} rM$, $M \xrightarrow{\beta} 0$, for which (cf. (4.13)-(4.14)) $G = (r, -1)$, $\rho_1^\sigma(k) = \alpha$, $\rho_2^\sigma(k) = \beta k$, $f^\sigma(k) = r\alpha - \beta k$. Since the

---

[28]Related to the CV, but not dimensionless, is the "Fano factor" defined as $\frac{\sigma^2(X)}{\mathbb{E}[X]}$.

reactions are of order ≤ 1, the FD formula is exact. We have that $J = -\beta$ and $B(\mu) = \alpha r^2 + \beta\mu$, so that:

$$\dot{\mu} \;=\; f(\mu) \;=\; \alpha r - \beta\mu \tag{4.54a}$$
$$\dot{\Sigma} \;=\; -2\beta\Sigma + B(\mu) \;=\; -2\beta\Sigma + \alpha r^2 + \beta\mu \,. \tag{4.54b}$$

In particular, at steady state we have:

$$\mu \;=\; \frac{\alpha r}{\beta} \;=\; \lambda r$$

$$\Sigma \;=\; \frac{\alpha r^2 + \beta\frac{\alpha r}{\beta}}{2\beta} \;=\; \frac{\alpha r^2 + \alpha r}{2\beta} \;=\; \lambda\frac{r(r+1)}{2}$$

where we again denote $\lambda = \frac{\alpha}{\beta}$. Thus,

$$\mathbf{cv}\,[M]^2 \;=\; \lambda\frac{r(r+1)}{2} \Big/ \lambda^2 r^2 \;=\; \frac{r+1}{2r}\frac{1}{\lambda}$$

which specializes to $1/\lambda$ in the Poisson case (no bursting, $r = 1$). Note that noise, as measured by the CV, is lower when r is higher, but never lower than $1/2$ of the Poisson rate.

This example is a typical one in which experimental measurement of means (or of the deterministic model) does not allow one to identify a parameter (r in this case), but the parameter can be identified from other statistical information: r (as well as λ) can be recovered from μ and Σ.

Next, we return to the dimerization example given by the reactions in (4.15), $0 \xrightarrow{\alpha} A$, $A + A \xrightarrow{\beta} 0$, for which (cf. (4.16)-(4.17)) $G = (1, -2)$, $\rho_1^\sigma(k) = \alpha$, $\rho_2^\sigma(k) = \frac{\beta k(k-1)}{2}$, $f^\sigma(k) = \alpha + \beta k - \beta k^2$. Some reactions are now of order 2, and the FD formula is not exact. In fact,

$$\dot{\mu} \;=\; \mathbb{E}\left[f^\sigma(X(t))\right] \;=\; \alpha + 2\beta m - \beta\mathbb{E}\left[X(t)^2\right] \;=\; \alpha + \beta\mu - \beta(\Sigma + \mu^2) \;=\; \alpha + \beta\mu - \beta\mu^2 - \beta\Sigma$$

shows that the mean depends on the variance

Exercise. Obtain an equation for $\dot{\Sigma}$ (which will depend on moments of order three).

Finally, we study in some detail the *transcription/translation* model (4.6)-(4.18):

$$0 \xrightarrow{\alpha} M \xrightarrow{\beta} 0 \,, \quad M \xrightarrow{\theta} M + P \,, \quad P \xrightarrow{\delta} 0 \,.$$
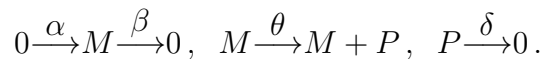
We had from (4.19)-(4.20) that

$$\Gamma \;=\; \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \,, \quad \rho_1^\sigma(k) = \alpha \,, \quad \rho_2^\sigma(k) = \beta k_1 \,, \quad \rho_3^\sigma(k) = \theta k_1 \,, \quad \rho_4^\sigma(k) = \delta k_2 \,.$$

and (writing "(M, P)" instead of $k = (k_1, k_2)$):

$$f^\sigma(M, P) = \begin{pmatrix} \alpha - \beta M \\ \theta M - \delta P \end{pmatrix} \,.$$

Since *all reactions are of order at most one, the FD formula is exact*. There are 5 differential equations: 2 for the means and 3 (omiting one by symmetry) for the covariances. For means we have:

$$\dot{\mu}_M \;=\; \alpha - \beta\mu_M \tag{4.55a}$$
$$\dot{\mu}_P \;=\; \theta\mu_M - \delta\mu_P \,. \tag{4.55b}$$

Now, using the formula $\mathbb{E}\left[B(X(t))\right] = \Gamma \operatorname{diag}\left(\mathbb{E}\left[\rho_1^\sigma(X(t))\right], \ldots, \mathbb{E}\left[\rho_m^\sigma(X(t))\right]\right)\Gamma'$ (see (4.39)) for the expectation of the diffusion term, we obtain that $B(\mu)$ equals:

$$
\begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \alpha & & & \\ & \beta\mu_M & & \\ & & \theta\mu_M & \\ & & & \delta\mu_P \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} \alpha + \beta\mu_M & 0 \\ 0 & \theta\mu_M + \delta\mu_P \end{pmatrix}.
$$

Also,

$$
J = \text{Jacobian of } \begin{pmatrix} \alpha - \beta M \\ \theta M - \delta P \end{pmatrix} = \begin{pmatrix} -\beta & 0 \\ \theta & -\delta \end{pmatrix}.
$$

It follows that the variance part of the FD equation

$$
\dot{\Sigma} = \Sigma J' + J\Sigma + B
$$

is (omitting the symmetric equation for Σ_{PM}):

$$
\begin{align}
\dot{\Sigma}_{MM} &= -2\beta\,\Sigma_{MM} + \alpha + \beta\mu_M \tag{4.56a}\\
\dot{\Sigma}_{PP} &= -2\delta\,\Sigma_{PP} + 2\,\theta\Sigma_{MP} + \theta\mu_M + \delta\mu_P \tag{4.56b}\\
\dot{\Sigma}_{MP} &= \theta\,\Sigma_{MM} - (\beta + \delta)\,\Sigma_{MP}. \tag{4.56c}
\end{align}
$$

In particular, at steady state we have the following mean number of proteins:

$$
\mu_P = \frac{\alpha\theta}{\beta\delta} \tag{4.57}
$$

and the following squared coefficient of variation for protein numbers:

$$
\mathbf{cv}\left[P\right]^2 = \frac{\Sigma_{PP}}{\mu_P^2} = \frac{(\theta + \beta + \delta)\beta\delta}{\alpha\theta(\beta + \delta)} = \frac{1}{\mu_P} + \frac{1}{\mu_M}\frac{\delta}{\beta + \delta}. \tag{4.58}
$$

Exercise. Prove the above formula for the CV. Show also that $\Sigma_{MP} = \frac{\theta\alpha}{\beta(\beta+\delta)}$.

The first term in (4.58) is usually referred to as the "*intrinsic noise*" of transcription, in the sense that this is what the cv would be, if M was constant (so that P would be a Poisson process).

The second is term is usually referred to as the "*extrinsic noise*" of transcription, due to mRNA variability.

Notice that the total noise is bounded from below by the intrinsic noise, and from above by the sum of the intrinsic noise and the mRNA noise, in the following sense:

$$
\frac{1}{\mu_P} \leq \mathbf{cv}\left[P\right]^2 \leq \frac{1}{\mu_P} + \frac{1}{\mu_M}
$$

(the second inequality because $\frac{\delta}{\beta+\delta} < 1$).

Also, note that even if the mean protein number $\mu_P \gg 1$, the second term, $\frac{1}{\mu_M}\frac{\delta}{\beta+\delta}$, may be large, so that extrinsic noise may dominate even in "large" systems.

Moreover, even accounting for much faster mRNA than protein degradation: $\beta \gg \delta$, which implies $\frac{\delta}{\beta+\delta} \ll 1$, this term may well be large if $\mu_M \ll 1$.

Yet another way to rewrite the total protein noise is as follows:

$$\mathbf{cv}\,[P]^2 \;=\; \frac{1}{\mu_P}\left[1+\frac{b}{1+\eta}\right]$$

where $\eta = \frac{\theta}{\beta}$ is the ratio of mRNA to protein lifetimes, and $b = \theta/\beta$ is the *burst factor* of the translation/transcription process. The number η is typically very small, in which case we have the approximation $\mathbf{cv}\,[P]^2 \approx \frac{1+b}{\mu_P}$. Since b is typically much larger than one, this means that the noise in P is much larger than would be expected for a Poisson random variable $(1/\mu_P)$.[29]

Exercise. Give an argument to justify why the burst factor may be thought of as the average number of proteins produced per transcript (i.e, during an mRNAs' lifetime). (The argument will be similar to the one used in the context of epidemics.)

---

[29] According to M. Thattai and A. Van Oudenaarden, "Intrinsic noise in gene regulatory networks," Proc. Natl Acad. Sci. USA 98, 8614-8619, 2001, which is one of the foundational papers in the field, 'typical values for b are 40 for lacZ and 5 for lacI'.

4.8   Conservation laws and stoichiometry

Suppose that $\nu \in \ker \Gamma'$, i.e. its transpose ν' is in the left nullspace of the stoichiometry matrix Γ, $\nu'\Gamma = 0$. For differential equation models of chemical reactions, described as $\dot{x} = \Gamma R(x)$, it is clear that $\nu'x(t)$ is constant, because $d(\nu'x)/dt = \nu'\Gamma R(x) = 0$. A similar invariance property holds for solutions of the CME. The basic observation is as follows.

Suppose that $\nu'\gamma_j = 0$ and that $c \in \mathbb{Z}$. Then

$$\sum_{\nu'k=c} \rho_j^\sigma(k - \gamma_j)\, p_{k-\gamma_j}(t) \;=\; \sum_{\nu'k=c} \rho_j^\sigma(k)\, p_k(t)\,,$$

where the sums are being taken over all those $k \in \mathbb{Z}_{\geq 0}^n$ such that $\nu'k = c$ (recall the convention that $\rho_j^\sigma(\ell) = 0$ if $\ell \notin \mathbb{Z}_{\geq 0}^n$). This is clear by a change of variables $\ell = k - \gamma_j$, since $\nu'k = c$ if and only if $\nu(k - \gamma_j) = 0$.

Therefore, for any $\nu \in \ker \Gamma'$ it follows that (dropping arguments t):

$$\frac{d}{dt}\sum_{\nu'k=c} \dot{p}_k \;=\; \sum_{j=1}^m \sum_{\nu'k=c} \left[\rho_j^\sigma(k - \gamma_j)\, p_{k-\gamma_j} - \rho_j^\sigma(k)\, p_k\right] \;=\; \sum_{j=1}^m 0 \;=\; 0\,.$$

So $\sum_{\nu'k=c} p_k(t)$ is constant.

Suppose that the initial state $X(0)$ is known to satisfy $\nu'X(0) = c$. In other words, $\sum_{\nu'k=c} p_k(0) = 1$. It then follows that $\sum_{\nu'k=c} p_k(t) = 1$, which means that, with probability one, $\nu'X(t) = c$ for each $t \geq 0$. This invariance property is an analogue of the one for deterministic systems.

The limit $\pi = p(\infty)$, if it exists, of the distribution vector satisfies the constraint $\sum_{\nu'k=c} \pi_k = 1$. This constraint depends on the initial conditions, through the number c. Steady-state solutions of the CME are highly non-unique when there are conservation laws. To deal with this problem, the usual approach consists of reducing the space by expressing redundant species in terms of a subset of "independent" species, as follows.

Consider a basis $\nu_1, \ldots, \nu_s$ of $\ker \Gamma'$. If it is known that $\nu_i'X(0) = c_i$ for $i = 1, ..., s$, then the above argument says that $\sum_{\nu_i'k=c_i} p_k(t) = 1$ and $\nu_i'X(t) = c_i$ for each $i = 1, ..., s$ and each $t \geq 0$. This fact typically allows one to reduce the Markov chain to a smaller subset.

The simplest example is that of the reaction network

$$A \xrightarrow{\mu} B\,, \quad B \xrightarrow{\nu} A\,,$$
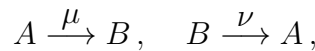
for which we have:

$$\Gamma \;=\; \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}\,, \quad \rho_1^\sigma(k) = \mu k_1\,, \quad \rho_2^\sigma(k) = \nu k_2\,.$$

We pick $s = 1$ and $\nu = (1,1)'$.

Suppose that, initially, there is just one unit of A, that is, $X(0) = (A(0), B(0))' = (1,0)'$. Thus $\nu'X(0) = 1$, from which it follows that $A(t) + B(t) = \nu'X(t) = 1$ for all $t \geq 0$ (with probability one), or equivalently, that $\sum_{k_1+k_2=1} p_k(t) = 1$ for all $t \geq 0$.

Since $p_k(t) = 0$ if either $k_1 < 0$ or $k_2 < 0$, this amounts to saying that $p_{(1,0)'}(t) + p_{(0,1)'}(t) = 1$ for all $t \geq 0$, and $p_k(t) = 0$ for all other k.

If we are only interested in the initial condition $X(0) = (1,0)'$, there is no need to compute $p_k(t)$ except for these two k's. The finite Markov chain with the two states $(1,0)'$ and $(0,1)'$ carries all the information that we care for. Moreover, since $p_{(0,1)'}(t) = 1 - p_{(1,0)'}(t)$, it is enough to consider the differential equation for $p(t) = p_{(1,0)'}(t)$:

$$
\dot{p} = \rho_1^\sigma \left(\begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} -1 \\ 1 \end{pmatrix} \right) p_{\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} -1 \\ 1 \end{pmatrix} \right)} + \rho_2^\sigma \left(\begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right) p_{\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right)}
$$
$$
- \rho_1^\sigma \begin{pmatrix} 1 \\ 0 \end{pmatrix} p_{\begin{pmatrix} 1 \\ 0 \end{pmatrix}} - \rho_2^\sigma \begin{pmatrix} 1 \\ 0 \end{pmatrix} p_{\begin{pmatrix} 1 \\ 0 \end{pmatrix}} \, .
$$

Since $\rho_1^\sigma(2,-1)' = \rho_2^\sigma(1,0)' = 0$, $\rho_2^\sigma(0,1)' = \nu$, and $\rho_1^\sigma(1,0)' = \mu$ and $p_{(0,1)'} = 1 - p$, we conclude that

$$
\dot{p} = (1-p)\nu - p\mu, \quad p(0) = 1,
$$

so

$$
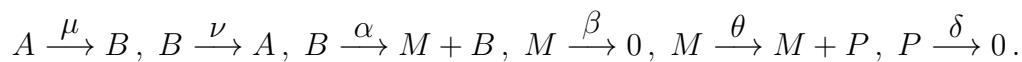p(t) = \frac{\nu}{\mu + \nu} + e^{-(\mu+\nu)t} \left(1 - \frac{\nu}{\mu + \nu} \right).
$$

In particular, at steady state,

$$
p_{\begin{pmatrix} 1 \\ 0 \end{pmatrix}}(\infty) = \frac{\nu}{\mu + \nu}, \quad p_{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}(\infty) = \frac{\mu}{\mu + \nu},
$$

i.e., the steady-state distribution is Bernoulli with parameter $\frac{\nu}{\mu+\nu}$.

Exercise. Suppose that, in the reaction network $A \xrightarrow{\mu} B$, $B \xrightarrow{\nu} A$, we know that initially, there are just r units of A, that is, $X(0) = (A(0), B(0))' = (r, 0)'$. Show how to reduce the CME to a Markov chain on $s + 1$ states, and that the steady-state probability distribution is a binomial distribution.

Exercise. The example of $A \xrightarrow{\mu} B$, $B \xrightarrow{\nu} A$ with $X(0) = (A(0), B(0))' = (r, 0)'$ can be thought of as follows: A is the inactive form of a gene, and B is its active form. There are a total of r copies of the same gene, and the activity of each switches randomly and independently. Suppose that we now consider transcription and translation, where transcription is only possible when one of these copies of the gene is active. This leads to the following system:

$$
A \xrightarrow{\mu} B, \ B \xrightarrow{\nu} A, \ B \xrightarrow{\alpha} M + B, \ M \xrightarrow{\beta} 0, \ M \xrightarrow{\theta} M + P, \ P \xrightarrow{\delta} 0.
$$

1. Write down the CME for this system.

2. Assuming only one copy of the gene, $r = 1$, compute (using the FD method or generating functions) the steady-state mean and standard deviation of M.

3. Optional (very tedious computation): again with $r = 1$, use the FD formula to compute the steady-state mean and standard deviation of P.

4. Optional: repeat the calculations with an arbitrary copy number r.

4.9 Relations to deterministic equations, and approximations

In this section, we briefly discuss various additional topics, in an informal fashion. All propensities are mass-action type now.

4.9.1 Deterministic chemical equations

The mean of the state $X(t)$ satisfies the differential equation (4.29): $\frac{d}{dt}\mathbb{E}\left[X(t)\right] = \mathbb{E}\left[f^\sigma(X(t))\right]$. This suggests the approximation

$$\frac{d}{dt}\mathbb{E}\left[X(t)\right] \approx f^\sigma(\mathbb{E}\left[X(t)\right]), \tag{4.59}$$

which is an equality when the reactions have order 0 or 1. This would also be an equality of the variability of $X(t)$ were small. However, in general, the variance of $X(t)$ is large, of the order of the volume Ω in which the reaction takes place, as we discuss later.

On the other hand, if we consider the *concentration* $Z(t) = X(t)/\Omega$, this quantity has variance of order $\Omega/\Omega^2 = 1/\Omega$. So, for concentrations, and assuming that Ω is large, it makes sense to expect that the analog of (4.59) will be very accurate.

Now, to get a well-defined meaning of concentrations $Z(t) = X(t)/\Omega$ as $\Omega \to \infty$, $X(t)$ must also be very large. (Since otherwise $Z(t) = X(t)/\Omega \approx 0$.) This is what one means by a "thermodynamic limit" in physics.

What equation is satisfied by $\mathbb{E}\left[Z(t)\right]$? To be precise, let us consider the stochastic process $Z(t) = \frac{X(t)}{\Omega}$ that describes concentrations as opposed to numbers of units. Equation (4.29) said that $\frac{d}{dt}\mathbb{E}\left[X(t)\right] = \mathbb{E}\left[f^\sigma(X(t))\right]$. Therefore,

$$\frac{d}{dt}\mathbb{E}\left[Z(t)\right] = \frac{1}{\Omega}\frac{d}{dt}\mathbb{E}\left[X(t)\right] = \frac{1}{\Omega}\mathbb{E}\left[f^\sigma(X(t))\right] = \mathbb{E}\left[\frac{1}{\Omega}f^\sigma(\Omega Z(t))\right]. \tag{4.60}$$

The numbers $Z(t)$, being concentrations, should be expected to satisfy some sort of equation that does not in any way involve volumes. Thus, we want to express the right-hand side of (4.60) in a way that does not involve Ω-dependent terms. Unfortunately, this is not possible without appealing to an approximation. To illustrate the problem, take a homodimerization reaction, which will contribute terms of the form $\frac{1}{\Omega}k(k-1)$ to the vector field f^σ. Then the right-hand side of (4.60) will involve an expression

$$\frac{1}{\Omega^2}\left(\Omega Z(t)\right)\left(\Omega Z(t) - 1\right) = \left(\frac{\Omega Z(t)}{\Omega}\right)\left(\frac{\Omega Z(t)}{\Omega}\right)\left(1 - \frac{1}{Z(t)\Omega}\right) = Z(t)^2\left(1 - \frac{1}{Z(t)\Omega}\right)$$

Thus, we need to have $Z(t)\,\Omega \gg 1$ in order to eliminate Ω-dependence. This is justified provided that $\Omega \to \infty$ and $Z(t) \not\to 0$. More generally, the discussion is as follows.

The right-hand side of (4.60) involves $\frac{1}{\Omega}f^\sigma$, which is built out of terms of the form $\frac{1}{\Omega}\rho_j^\sigma$, where the propensities for mass-action kinetics are $\rho_j^\sigma(k) = \frac{c_j}{\Omega^{A_j-1}}\binom{k}{a_j}$ for each $j \in \{1, \ldots, m\}$.

The combinatorial numbers $\binom{k}{a_j} = \prod_{i=1}^n \binom{k_i}{a_{ij}}$ can be approximated as follows. For each $j \in \{1, \ldots, m\}$, using the notation $a_j! = \prod_{i=1}^n a_{ij}!$, we have:

$$\binom{k}{a_j} = \frac{k^{a_j}}{a_j!}\left[1 + O\left(\frac{1}{k}\right)\right]. \tag{4.61}$$

For example, since

$$
\begin{pmatrix} k_1 \\ 3 \end{pmatrix} = \frac{1}{3!} k_1(k_1 - 1)(k_1 - 2) = \frac{k_1^3}{3!} \left[1 + \frac{1}{k_1} P \left(\frac{1}{k_1} \right) \right]
$$

$$
\begin{pmatrix} k_2 \\ 2! \end{pmatrix} = \frac{1}{2} k_2(k_2 - 1) = \frac{k_2^2}{2!} \left[1 + \frac{1}{k_2} Q \right]
$$

with $P(x) = -3 + 2x$, and $Q = -1$, then if $n = 2$ and $a_j = (3, 2)'$ (that is, the reaction $\mathcal{R}_j$ consumes three units of S_1 and two of S_2), we have that

$$
\begin{pmatrix} k_1 \\ 3 \end{pmatrix} \times \begin{pmatrix} k_2 \\ 2 \end{pmatrix} = \frac{k_1^3 k_2^2}{3!2!} \left[1 + \frac{1}{k_1} P + \frac{1}{k_2} Q + \frac{1}{k_1 k_2} PQ \right].
$$

Let us introduce the following functions:

$$
\rho_j^c(s) = \frac{c_j}{a_j!} s^{a_j}
$$

where, for each $s \in \mathbb{R}_{\geq 0}^n$ with components s_i,

$$
s^{a_j} = \prod_{i=1}^{n} s_i^{a_{ij}}
$$

(with the convention that $s^0 = 1$ for all s).

Observe that, with our notations,

$$
\frac{k^{a_j}}{\Omega_j^A} = \left(\frac{k}{\Omega} \right)^{a_j}.
$$

So, we consider the approximation:

$$
\frac{1}{\Omega} \rho_j^\sigma(k) = \frac{c_j}{\Omega^{A_j}} \begin{pmatrix} k \\ a_j \end{pmatrix} = \frac{c_j}{\Omega^{A_j}} \frac{k^{a_j}}{a_j!} \left[1 + O \left(\frac{1}{k} \right) \right] = \rho_j^c \left(\frac{k}{\Omega} \right) \left[1 + O \left(\frac{1}{k} \right) \right] \approx \rho_j^c \left(\frac{k}{\Omega} \right),
$$
(4.62)

which is valid if

both $k \to \infty$ and $\Omega \to \infty$ in such a way that the ratio k/Ω remains constant.

This type of limit is often referred to as a "thermodynamic limit". It is interpreted as saying that both the copy numbers and volume are large, but the *concentrations* or *densities* are not. Another way to think of this is by thinking of a larger and larger volume in which a population of particles remains at constant density (so that the number of particles scales like the volume). For purposes of this discussion, let us just agree to say that "in the thermodynamic approximation" will mean whenever the approximation has been performed.

Recall from Equation (4.30) that $f^\sigma(k) = \Gamma R^\sigma(k)$, where $R^\sigma(k) = (\rho_1^\sigma(k), \ldots, \rho_m^\sigma(k))'$ and Γ is the stoichiometry matrix. Let $R(x)$ be defined for any non-negative real vector s as follows:

$$
R(x) := (\rho_1^c(s), \ldots, \rho_m^c(s))'
$$
(4.63)

and let

$$
f(s) := \Gamma R(s).
$$
(4.64)

Under the thermodynamic approximation for (4.62),

$$\frac{1}{\Omega} f^\sigma(k) \approx f\left(\frac{k}{\Omega}\right).$$

That is to say, (4.60) becomes

$$\frac{d}{dt} \mathbb{E}\left[Z(t)\right] \approx \mathbb{E}\left[f(Z(t))\right]. \tag{4.65}$$

We achieved our goal of writing an (approximate) expression that is volume-independent, for the rate of change of the mean concentration

Provided that the variance of $Z(t)$ is small compared to its mean, then we may approximate $\mathbb{E}\left[f(Z(t))\right] \approx f(\mathbb{E}\left[Z(t)\right])$ and write

$$\frac{d}{dt} \mathbb{E}\left[Z(t)\right] \approx f(\mathbb{E}\left[Z(t)\right]).$$

This argument motivates the form of the *deterministic chemical reaction equation*[30] which is (using dot for time derivative, and omitting the time argument):

$$\dot{s} = f(s) = \Gamma R(s). \tag{4.66}$$

Observe that we may also write this deterministic equation as an equation on the abundances $x(t)$ of the species, where $x(t) = \Omega s(t)$. The equation is:

$$\dot{x} = f^\#(x) = \Gamma R^\#(x) \tag{4.67}$$

where

$$R^\#(x) := \left(\rho_1^\#(x), \ldots, \rho_m^\#(x)\right)' \tag{4.68}$$

and

$$\rho_j^\#(x) = \Omega \rho_j^c\left(\frac{x}{\Omega}\right) = \frac{c_j}{\Omega^{A_j-1}} \frac{x^{a_j}}{a_j!}.$$

The only difference with the expression for concentrations is that now there is a denominator which depends on volume.

Both forms of deterministic equations are used in the literature, usually not distinguishing among them. *They both may be written in the same form*, using rates "$\rho(u) = k_j u^{a_j}$" after collecting all constants into k_j, and the only difference is the expression of k_j in terms of the volume. For problems in which the deterministic description is used, and if one is not interested in the stochastic origin of the reaction constants k_j, this is all unimportant. In fact, in practice the coefficients k_j are often estimated by fitting to experimental data, using a least-squares or maximum-likelihood method. In that context, *the physical origin of the coefficients, and their volume dependence or lack thereof, plays no role.*

4.9.2   Unit Poisson representation

We next discuss an integral representation which is extremely useful in the theoretical as well as in the intuitive understanding of the behavior of the process $X(t)$.

---

[30]Also called a "mean field equation" in physics

To motivate this representation, note first that a vector $x(t)$ is a solution of the deterministic differential equation $\dot{x}(t) = f(x(t))$ with initial condition $x(0) = x_0$ if and only if $x(t) = x_0 + \int_0^t f(x(\tau))\, d\tau$ for all t. This reformulation as an integral equation is merely a statement of the Fundamental Theorem of Calculus, and in fact is a key step in proving existence theorems for differential equations (by looking for fixed points of the operator $x \mapsto x_0 + \int_0^t f(x(\tau))\, d\tau$ in function space).

Specialized to the chemical reaction case, using abundances x, $f(x) = f^{\#}(x) = \Gamma\, R^{\#}(x)$, the integral equation reads:

$$x(t) \;=\; = \; x(0) + \sum_{j=1}^m \gamma_j\, y_j(t)\,, \qquad \text{where } y_j(t) = \int_0^t \rho_j^{\#}(x(\tau))\, d\tau\,. \tag{4.69}$$

The quantity $y_j(t)$ may be thought of as the number of reactions of $\mathcal{R}_j$ that have taken place until time t, because each such reaction adds γ_j to the state. As $\dot{y}_j(t) = \rho_j^{\#}(x(t))$, $\rho_j^{\#}$ can be interpreted as the rate at which the reaction $\mathcal{R}_j$ takes place.

We now turn to the stochastic model. The random state $X(t)$ at time t is obtained from a sequence of jumps:

$$X(t) \;=\; X(0) + W_1 + \ldots + W_N\,.$$

Collecting all the terms W_v that correspond to events in which $\mathcal{R}_j$ fired, and keeping in mind that, every time that the reaction $\mathcal{R}_j$ fires, the state changes by $+\gamma_j$, there results:

$$X(t) \;=\; X(0) + \sum_{j=1}^m \gamma_j\, \widetilde{Y}_j(t)\,, \tag{4.70}$$

where $\widetilde{Y}_j$ counts *how many times* the reaction j has taken place from time 0 until time t. The stochastic Equation (4.70) is a counterpart of the deterministic Equation (4.69). Of course, $\widetilde{Y}_j(t)$ depends on the past history $X(\tau)$, $\tau < t$. The following *Poisson representation* makes that dependence explicit:

$$X(t) \;=\; X(0) + \sum_{j=1}^m \gamma_j\, Y_j\left(\int_0^t \rho_j^{\sigma}(X(\tau))\, d\tau \right)\,, \tag{4.71}$$

where the Y_j's are m independent and identically distributed ("IID") Poisson processes with unit rate. This most beautiful formula is exact and requires no approximations[31]. Here we simply provide an intuitive idea of why one may expect such a formula to hold. The intuitive idea is based on an argument as the one used to derive the SSA.

If $k = X_{v-1}(t_{v-1})$ is the state right after the $(v-1)$st jump, then the *time* until the next jump is given by the variable $\mathcal{T}_k$, which is exponential with the parameter in Equation (4.27), $\lambda_k = \sum_{j=1}^m \rho_j^{\sigma}(k)$. If the state k does not change much, then these distributions do not depend strongly on k, and we can say that reactions occur at times that are separated by an exponentially distributed random variable $\mathcal{T}$ with rate λ. From basic probability theory, we know that this means that the total number of reactions during an interval of length t is Poisson distributed with parameter $t\lambda$. That is to say, there is a Poisson process Y with rate λ that counts how many reactions happen in any given interval.

---

[31]For details, including proofs, see S.N. Ethier and T.G. Kurtz, *Markov processes: Characterization and convergence*, John Wiley & Sons, New York, 1986.

The random choice of *which* reaction takes place is distributed according to the probabilities

$$\mathbb{P}\left[\text{next reaction is } \mathcal{R}_j\right] \;=\; d_j^{(k)} \;=\; \frac{\rho_j^\sigma(k)}{\sum_{j=1}^m \rho_j^\sigma(k)} \,.$$

If the reaction events form a Poisson process with parameter λ, and if at each time the reaction to be used is picked according to a discrete distribution with $p_j = d_j = \rho_j^\sigma/\lambda$ (we drop "k" since we are assuming that it is approximately constant), then the events "$\mathcal{R}_j$ fires" form a "thinning" of the Poisson stream and hence are known, again from elementary probability theory, to be themselves Poisson distributed, with parameter $d_j\lambda = \rho_j^\sigma$.

This means, putting back the k now, that the number of reactions of type $\mathcal{R}_j$ that occur are distributed according to $\widetilde{Y}_j(t) = Y_j(\rho_j^\sigma(k)t)$, where the Y_j are independent unit Poisson processes[32] (independence also assumes that k is approximately constant during the interval). Now, if we break up a long interval into small intervals of length dt, in each of which we assume that k is constant (somewhat analogous to making an approximation of an integral using a rectangle rule), we have that the total $\widetilde{Y}_j(t)$ is a sum of Poisson random variables, one for each sub-interval, with rates $\rho_j^\sigma(k)dt$. A sum of (independent) Poisson random variables with rates $\mu_1, \dots, \mu_\nu$ is Poisson with rate $\mu_1 + \dots + \mu_\nu$, and, as the intervals get smaller, this sum approximates the integral $\int_0^t \rho_j^\sigma(X(\tau))\, d\tau$, if the $\mu_i = \rho_j^\sigma(X(\tau_i))$. This results in the formula (4.71), though of course the argument is not at all rigorous as given.

4.9.3 Diffusion approximation

A *stochastic differential equation* (SDE) is an ordinary differential equation with noise terms in its right-hand side, so that its solution is random.[33] The Markov jump process $X(t)$ is not the solution of an SDE, since by definition, it is discrete-valued.[34] However, there is an SDE whose solutions give a so-called *diffusion approximation* of $X(t)$.[35] The diffusion approximation is useful when numbers of species are "large enough". (But not so large that the equation becomes basically deterministic and so there is no need for stochastics to start with.) It arises as a normal approximation of a Poisson process. We very roughly outline the construction, as follows.

We consider the formula (4.71), which works on any interval $[t, t + h]$:

$$X(t + h) \;=\; X(t) + \sum_{j=1}^m \gamma_j\, Y_j\left(\int_t^{t+h} \rho_j^\sigma(X(\tau))\, d\tau\right)$$

where the Y_j's are IID unit Poisson random processes.

In general, under appropriate conditions ($\lambda \gg 1$), if a variable Y is Poisson with parameter λ, then it is well approximated by a normal random variable N with mean λ and variance λ (this is a special

---

[32]Saying that Z is a Poisson process with rate λ is the same as saying that $Z(t) = Y(\lambda t)$, where Y is a unit-rate Poisson process.

[33]In physics, SDE's are called *Langevin equations*.

[34]Of course, there is an ODE associated to $X(t)$, namely the CME. But the CME is a *deterministic* differential equation for the *probability distribution* of $X(t)$, not for the sample paths of $X(t)$.

[35]As if things were not confusing enough already, there is yet another (deterministic) differential equation that enters the picture, namely the *Fokker-Planck Equation* (FPE), which, describes the evolution of the probability distribution of the state of the SDE, just like the CME describes the evolution of the probability distribution of the state $X(t)$. The FPE is a PDE (enough acronyms?), because the probability the state of the SDE is a continuous variable, hence requiring a variable for space as well as time.

case of the Central Limit Theorem). Equivalently,

$$Y \approx \lambda + \sqrt{\lambda}\, N_0\,,$$

where N_0 is an $\mathcal{N}(0,1)$ random variable.

We make this approximation in the above formula. We denote the random variable N_0 as "$N_j(t)$" to indicate the fact that we have a different one for each j and for each interval $[t, t+h]$ where the approximation is made. Note that, given the initial state $X(t)$, the changes in the interval $[t, t+h]$ are independent of changes in previous intervals; thus the $N_j(t)$ are independent of previous values. Using that $f = \sum_j \gamma_j \rho_j^\sigma$:

$$
\begin{aligned}
X(t+h) &\approx X(t) + \sum_{j=1}^m \gamma_j \left[\left(\int_t^{t+h} \rho_j^\sigma(X(\tau))\, d\tau \right) + \sqrt{ \left(\int_t^{t+h} \rho_j^\sigma(X(\tau))\, d\tau \right) }\, N_j(t) \right] \\
&\approx X(t) + f(X(t))h + \sum_{j=1}^m \gamma_j \sqrt{\rho_j^\sigma(X(t))}\, \sqrt{h}\, N_j(t)\,.
\end{aligned}
$$

The expressions $\sqrt{h}\, N_j(t)$ correspond to increments on time h of a Brownian motion. Thus (dividing by h and letting $h \to 0$), formally we obtain:

$$dX(t) \approx f(X(t))\, dt + \sum_{j=1}^m \gamma_j \sqrt{\rho_j^\sigma(X(t))}\, B_j(t)$$

where the B_t are independent standard Brownian motion processes.[36]

4.9.4   Relation to deterministic equation

We next sketch why, in the thermodynamic limit, the solution $s(t)$ of the deterministic equation for concentrations provides a good approximation of the mean $\mathbb{E}\left[X(t)\right]$.

We consider a thermodynamic limit, and let $Z(t) = X(t)/\Omega$. Then:

$$X(t) = X(0) + \sum_{j=1}^m \gamma_j Y_j \left(\int_0^t \rho_j^\sigma(\Omega Z(\tau))\, ds \right) = X(0) + \sum_{j=1}^m \gamma_j Y_j \left(\Omega \int_0^t \rho_j^c(Z(\tau))\, ds \right)\,.$$

On any fixed time interval, $Z(\tau)$ is bounded (assuming that there is a well-defined behavior for the densities in the thermodynamic limit), so that the variance of each $Y_j(\ldots)$ is $O(\Omega t)$ (if Y is a unit random process, the variance of $Y(\lambda t)$ is λt), and hence so is the variance of $X(t)$. On a bounded time interval, we may drop the "t" and just say that $\mathbb{V}\mathrm{ar}\left[X(t)\right] = O(\Omega)$

Now,

$$\frac{d}{dt}\mathbb{E}\left[Z(t)\right] = \frac{d}{dt}\mathbb{E}\left[X(t)/\Omega\right] = \frac{1}{\Omega}\frac{d}{dt}\mathbb{E}\left[X(t)\right] = \frac{1}{\Omega}f^\sigma(\mathbb{E}\left[X(t)\right]) + \frac{M}{\Omega} \approx f(\mathbb{E}\left[Z(t)\right]) + \frac{M}{\Omega}$$

---

[36]For technical reasons, one does not write the derivative form of the equation. The problem is that dB_j/dt is not well-defined as a function because B is highly irregular.

where "M" represents terms that involve central moments of $X(t)$ of order ≥ 2 (recall (4.34)). Moreover, M comes from a Taylor expansion of f^σ, and the nonlinear terms in f^σ (corresponding to all the reactions of order > 1) all have at least a factor $1/\Omega$. Thus, M is of order $O((1/\Omega) \times \mathbb{V}\mathrm{ar}\,[X(t)])$. Since, by the previous discussion, $\mathbb{V}\mathrm{ar}\,[X(t)] = O(\Omega)$, it follows that $M = O(1)$. We conclude that, in the thermodynamic limit,

$$\frac{d}{dt}\mathbb{E}\,[Z(t)] \;\approx\; f(\mathbb{E}\,[Z(t)]) \;+\; O\left(\frac{1}{\Omega}\right) \;\approx\; f(\mathbb{E}\,[Z(t)]),$$

which is (with equality) the deterministic equation.

Note, also, that $\mathbb{V}\mathrm{ar}\,[Z(t)] = \frac{1}{\Omega^2}\mathbb{V}\mathrm{ar}\,[X(t)] = O(1/\Omega)$. In other words, the "noise" in concentrations, as measured by their standard deviations, scales as $1/\sqrt{\Omega}$.

We close this section with a citation to a precise theorem of Kurtz[37] that provides one rigorous version of the above arguments. It says roughly that, on each finite time interval $[0, T]$, and for every $\varepsilon > 0$,

$$\mathbb{P}\,[\forall\, 0 \leq t \leq T,\; |Z(t) - s(t)| < \varepsilon] \;\approx\; 1$$

if Ω is large, where $Z = X/\Omega$ and $s(t)$ is the solution of the deterministic equation, assuming that $X(0) = s(0)$ (deterministic initial condition) and that the solution of $s(t)$ exists on this interval. In other words, "almost surely" the sample paths of the process, normalized to concentrations, are almost identical to the solution of the deterministic system. Of course, $X(0) = s(0)$ means $Z(0) = \Omega s(0)$, which makes no sense as $\Omega \to \infty$. So the precise statement is as follows:

Suppose that $X_\Omega(t)$ is a sample path of the process with volume Ω (that is, this is the volume that appears in the propensities), for each Ω. If $\frac{1}{\Omega}X_V(0) \to s(0)$, then:

$$\lim_{\Omega \to \infty} \mathbb{P}\left[\sup_{0 \leq t \leq T}\left|\frac{1}{\Omega}X_V(t) - s(t)\right| \geq \varepsilon\right] = 0$$

for all $T \geq 0$ and all $\varepsilon > 0$.

It is important to realize that, on longer time intervals (or we want a smaller error ε, the required Ω might need to be larger.

---

[37] See the previously cited book. Originally from T.G. Kurtz, "The relationship between stochastic and deterministic models for chemical reactions," *The Journal of Chemical Physics* **57**(1972): 2976-2978.

4.10   Problems for stochastic kinetics

1. Suppose that $p(t)$ satisfies the CME. Show that if $\sum_{k \in \mathbb{Z}_{\geq 0}^n} p_k(0) = 1$ then $\sum_{k \in \mathbb{Z}_{\geq 0}^n} p_k(t) = 1$ for all $t \geq 0$. (Hint: first, using that $\rho_j^\sigma(k - \gamma_j) = 0$ unless $k \geq \gamma_j$, observe that, for each $j \in \{1, \ldots, m\}$:

$$\sum_{k \in \mathbb{Z}_{\geq 0}^n} \rho_j^\sigma(k - \gamma_j) p_{k-\gamma_j} = \sum_{k \in \mathbb{Z}_{\geq 0}^n} \rho_j^\sigma(k) p_k$$

and use this to conclude that $\sum_{k \in \mathbb{Z}_{\geq 0}^n} p_k(t)$ must be constant. You may use without proof that the derivative of $\sum_{k \in \mathbb{Z}_{\geq 0}^n} p_k(t)$ with respect to time is obtained by term-by-term differentiation.)

2. Show, using induction on k, that, as claimed in the notes, $\pi_k = e^{-\lambda} \frac{\lambda^k}{k!}$, where $\lambda = \frac{\alpha}{\beta}$, solves

$$\alpha \pi_{k-1} + (k+1)\beta \pi_{k+1} - \alpha \pi_k - k\beta \pi_k = 0, \quad k = 0, 1, 2, \ldots$$

(the first term is not there if $k = 0$).

3. Write the CME for the bursting model.

4. Write the CME for the dimerization model.

5. Write the CME for the transcription/translation model. (Remember that now "k" is a vector (k_1, k_2).)

6. This is a problem regarding the SSA.

   Implement the SSA in your favorite programming system (MATLAB, Maple, Mathematica).

   (b) Take the mRNA/protein model described in the notes, pick some parameters, and an initial state; now plot many sample paths, averaging to get means and variances as a function of time, as well as steady state means and variances.

   (c) Compare the latter with the numbers obtained by using theory as described in the notes.

7. Show that an alternative way of writing the diffusion term in the FD equation is as follows:

$$\Gamma \operatorname{diag}\left(\mathbb{E}\left[\rho_1^\sigma(X(t))\right], \ldots, \mathbb{E}\left[\rho_m^\sigma(X(t))\right]\right) \Gamma'$$

(where "$\operatorname{diag}(r_1, \ldots, r_m)$" means a diagonal matrix with entries r_i in the diagonal).

8. Prove that, for the probability generating function P:

$$\left.\frac{\partial^2 P(z,t)}{\partial z_i \partial z_j}\right|_{z=1} = \begin{cases} \mathbb{E}\left[X_i(t)X_i(t)\right] & \text{if } i \neq j \\ \mathbb{E}\left[X_i(t)^2\right] - \mathbb{E}\left[X_i(t)\right] & \text{if } i = j. \end{cases}$$

9. For the mRNA example, derive the variance equation from the probability generating function, and show that the same result is obtained as in the notes.

10. For the mRNA example, solve explicitly the FD differential equations shown in the notes. (You may use matrix exponentials and variation of parameters, Laplace transforms, or whatever method you prefer.)
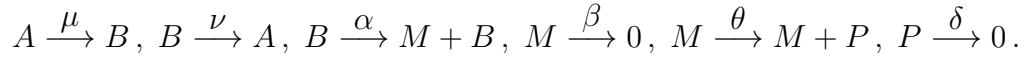
11. For the dimerization example, obtain an equation for $\dot{\Sigma}$ (which will depend on moments of order three).

12. For the transcription/translation example:

    (a) prove this formula for the squared coefficient of variation for protein numbers:

    $$\mathbf{cv}\,[P]^2 \;=\; \frac{\Sigma_{PP}}{\mu_P^2} \;=\; \frac{(\theta + \beta + \delta)\beta\delta}{\alpha\theta(\beta + \delta)} \;=\; \frac{1}{\mu_P} + \frac{1}{\mu_M}\frac{\delta}{\beta + \delta}\,.$$

    (b) Show that $\Sigma_{MP} = \frac{\theta\alpha}{\beta(\beta+\delta)}$.

13. Suppose that, in the reaction network $A\xrightarrow{\mu}B$, $B\xrightarrow{\nu}A$, we know that initially, there are just r units of A, that is, $X(0) = (A(0), B(0))' = (r, 0)'$. Show how to reduce the CME to a Markov chain on $s+1$ states, and that the steady-state probability distribution is a binomial distribution.

14. The example of $A\xrightarrow{\mu}B$, $B\xrightarrow{\nu}A$ with $X(0) = (A(0), B(0))' = (r, 0)'$ can be thought of as follows: A is the inactive form of a gene, and B is its active form. There are a total of r copies of the same gene, and the activity of each switches randomly and independently. Suppose that we now consider transcription and translation, where transcription is only possible when one of these copies of the gene is active. This leads to the following system:

    $$A \xrightarrow{\mu} B\,,\; B \xrightarrow{\nu} A\,,\; B \xrightarrow{\alpha} M + B\,,\; M \xrightarrow{\beta} 0\,,\; M \xrightarrow{\theta} M + P\,,\; P \xrightarrow{\delta} 0\,.$$

    (a) Write down the CME for this system.

    (b) Assuming only one copy of the gene, $r = 1$, compute (using the FD method or generating functions) the steady-state mean and standard deviation of M.

    (c) Optional (very tedious computation): again with $r = 1$, use the FD formula to compute the steady-state mean and standard deviation of P.

    (d) Optional: repeat the calculations with an arbitrary copy number r.

Appendix A

Review of ordinary differential equations

A.1 Modeling

A *differential equation* is just an equation which involves "differentials", that is to say, derivatives. A simple example is:

$$\frac{dy}{dt} = 0,$$

where we understand that y is a function of an independent variable t. (We use t because in many examples the independent variable happens to be time, but of course any other variable could be used. It is sometimes convenient to use informal notation, and refer to this example as "$y' = 0$" or as "$\dot{y} = 0$" (the latter notation is favored in engineering and applied mathematics), though such a notation blurs the distinction between *functions* and the *expressions* used to define them.

If $y' = 0$, y must be constant. In other words, the general solution of the given equation is $y \equiv c$, for some constant c.

Another easy example of a differential equation is:

$$\frac{dy}{dt} = -27.$$

This means that $y = y(t)$ has a graph which is a line with slope -27. The general solution of this equation is $y = -27t + c$, for some constant c.

An *initial value problem* is a problem in which we give a differential equation together with an extra condition at a point, like:

$$\frac{dy}{dt} = -27, \quad y(0) = 3.$$

There is a unique solution of this initial-value problem, namely $y(t) = -27t + 3$. It can be found by first finding the general solution $y = -27t + c$ and then plugging in $t = 0$ to get $3 = -27(0) + c$, so $c = 3$. This "initial" condition may be specified, of course, at any value of the independent variable t, (not just $t = 0$) for example:

$$\frac{dy}{dt} = -27, \quad y(2) = 3.$$

The solution of this initial-value problem can be also obtained by plugging into the general form $y = -27t + c$: we substitute $3 = y(2) = -27(2) + c$, which gives that $c = 57$, and so the solution is

$y(t) = -27t + 57$. Although the word "initial" suggests that we intend to start at that point and move forward in time, the solutions we have found are defined for all values of t. We will not always be so fortunate, but do we expect solutions defined on an interval with the "initial" value in the interior.

A slightly more complicated example of a differential equation is:

$$\frac{dy}{dt} = \sin t + t^2.$$

The general solution is (by taking antiderivatives) $y = -\cos t + t^3/3 + c$. Another example:

$$\frac{dy}{dt} = e^{-t^2}.$$

This equation has a general solution, but it cannot be expressed in terms of elementary functions like polynomials, trigs, logs, and exponentials. (The solution is the "error function" that is used in statistics to define the cumulative probability of a Gaussian or normal probability density.) One of the unfortunate facts about differential equations is that we cannot always find solutions as explicit combinations of elementary functions. So, in general, we have to use numerical, geometric, and graphical techniques in the analysis of properties of solutions.

The examples just given are too easy (even if $y' = e^{-t^2}$ doesn't look *that* easy), in the sense that they can all be solved, at least theoretically, by taking antiderivatives. The subject of differential equations deals with far more general situations, in which the unknown function y appears on both sides of the equation:

$$y' = f(t, y)$$

or even much more general types: systems of many simultaneous equations, higher order derivatives, and even partial derivatives when there are other independent variables (which leads to "partial differential equations" and are the subject of more advanced courses).

One aspect of differential equations is comparatively easy: if someone gives us an alleged solution of an equation, we can *check* whether this is so. Checking is much easier than finding! (Analogy: if I ask you to find a solution of the algebraic equation $10000x^5 - 90000x^4 + 65100x^3 + 61460x^2 + 13812x + 972 = 0$ it may take you some time to find one. On the other hand, if I tell you that $x = 3/2$ is a root, you can check whether I am telling the truth or not very easily: just plug in and see if you get zero.) For example, if someone claims that the function $y = 1 / (1 + t^2)$ is a solution of the equation $y' = -2ty^2$, we can check that she is right by plugging in:

$$\left(\frac{1}{1+t^2}\right)' = -\frac{2t}{(1+t^2)^2} = -2t\left(\frac{1}{1+t^2}\right)^2.$$

But if someone claims that $y = 1 / (1 + t)$ is a solution, we can prove him to be wrong:

$$\left(\frac{1}{1+t}\right)' = -\frac{1}{(1+t)^2} \neq -2t\left(\frac{1}{1+t}\right)^2$$

because the two last functions of t are not the same. They even have different values at $t = 0$.

About Modeling

Most applications of mathematics, and in particular, of differential equations, proceed as follows.

Starting from a "word problem" description of some observed behavior or characteristic of the real world, we attempt to formulate the simplest set of mathematical equations which capture the essential aspects. This set of equations represents a *mathematical model* of reality. The study of the model is then carried out using mathematical tools. The power of mathematics is that it allows us to make quantitative and/or qualitative conclusions, and predictions about behaviors which may not have been an explicit part of the original word description, but which nonetheless follow logically from the model.

Sometimes, it may happen the results of the mathematical study of the model turn out to be inconsistent with features found in the "real world" original problem. If this happens, we must modify and adapt the model, for example by adding extra terms, or changing the functions that we use, in order to obtain a better match. Good modeling, especially in science and engineering, is often the result of several iterations of the "model/reality-check/model" loop!

Unrestricted Population Growth

When dealing with the growth of a bacterial culture in a Petri dish, a tumor in an animal, or even an entire population of individuals of a given species, biologists often base their models on the following simple rule:

The increase in population during a small time interval of length Δt is proportional to Δt and to the size of the population at the start of the interval.

For example, statistically speaking, we might expect that one child will be born in any given year for each 100 people. The proportionality rule then says that two children per year are born for every 200 people, or that three children are born for each 100 people over three consecutive years. (To be more precise, the rate of increase should be thought of as the "net" rate, after subtracting population decreases. Indeed, the decreases may also assumed proportional to population, allowing the two effects to be combined easily.)

The rule is only valid for small intervals (small Δt), since for large Δt one should also include compounding effects (children of the children), just as the interest which a bank gives us on savings (or charges us on loan balances) gets compounded, giving a higher effective rate.

Let us call $P(t)$ the number of individuals in the population at any given time t. The simplest way to translate into math the assumption that "the increase in population $P(t + \Delta t) - P(t)$ is proportional to Δt and to $P(t)$" is to write

$$P(t + \Delta t) - P(t) = kP(t)\Delta t \tag{A.1}$$

for some constant k. Notice how this equation says that the increase $P(t + \Delta t) - P(t)$ is twice as big if Δt is twice as big, or if the initial population $P(t)$ is twice as big.

Example: in the "one child per 100 people per year" rule, we would take $k = 10^{-2}$ if we are measuring the time t in years. So, if at the start of 1999 we have a population of 100,000,000, then at the beginning of the year 2001 = 1999+2 the population should be (use $\Delta t = 2$):

$$P(2001) = P(1999) + 10^{-2}P(1999)\Delta t = 10^8 + 10^{-2}10^8(2) = 102,000,000$$

according to the formula. On the other hand, by the end of January 3rd, 1999, that is, with $\Delta t = 3/365$, we would estimate $P(1999 + 3/365) = 10^8 + 10^{-2}10^8(3/365) \approx 100,008,219$ individuals. Of course, there will be random variations, but on average, such formulas turn out to work quite well.

The equation (A.1) can only be accurate if Δt is small, since it does not allow for the "compound interest" effect. On the other hand, one can view (A.1) as specifying a step-by-step *difference equation* as follows. Pick a "small" Δt, let us say $\Delta t = 1$, and consider the following recursion:

$$P(t+1) = P(t) + kP(t) = (1+k)P(t) \tag{A.2}$$

for $t = 0, 1, 2, \ldots$. Then we compute $P(2)$ not as $P(0) + 2kP(0)$, but recursively applying the rule: $P(2) = (1+k)P(1) = (1+k)^2 P(0)$. This allows us to incorporate the compounding effect. It has the disadvantage that we cannot talk about $P(t)$ for fractional t, but we could avoid that problem by picking a smaller scale for time (for example, days). A more serious disadvantage is that it is hard to study difference equations using the powerful techniques from calculus. Calculus deals with things such as rates of change (derivatives) much better than with finite increments. Therefore, what we will do next is to show how the problem can be reformulated in terms of a differential equation. This is not to say that difference equations are not interesting, however. It is just that differential equations can be more easily studied mathematically.

If you think about it, you have seen many good examples of the fact that using derivatives and calculus is useful even for problems that seem not to involve derivatives. For example, if you want to find an integer t such that $t^2 - 189t + 17$ is as small as possible, you could try enumerating all possible integers (!), or you could instead pretend that t is a real number and minimize $t^2 - 189t + 17$ by setting the derivative to zero: $2t - 189 = 0$ and easily finding the answer $t = 94.5$, which then leads you, since you wanted an integer, to $t = 94$ or $t = 95$.

Back to our population problem, in order to use calculus, we must allow P to be any real number (even though, in population studies, only integers P would make sense), and we must also allow the time t to be any real number. Let us see where equation (A.1) leads us. If we divide by Δt, we have

$$\frac{P(t + \Delta t) - P(t)}{\Delta t} = kP(t).$$

This equation holds for small Δt, so we may let $\Delta t \to 0$. What is the limit of $\left(P(t + \Delta t) - P(t)\right) \big/ \Delta t$ as $\Delta t \to 0$? It is, as you remember from Calculus I (yes, you do), the derivative of P evaluated at t. So we end up with our first differential equation:

$$P'(t) = kP(t). \tag{A.3}$$

This is the differential equation for population growth. We may read it like this:

The rate of change of P is proportional to P.

The solution of this differential equation is easy: since $P'(t)/P(t) = k$, the chain rule tells us that

$$(\ln P(t))' = k,$$

and so we conclude that $\ln P(t) = kt + c$ for some constant c. Taking exponentials of both sides, we deduce that $P(t) = e^{kt+c} = Ce^{kt}$, where C is the new constant e^c. Evaluating at $t = 0$ we have that $P(0) = Ce^0 = C$, and we therefore conclude:

$$P(t) = P(0)e^{kt}.$$

(Actually, we cheated a little, because P'/P doesn't make sense if $P = 0$, and also because if P is negative then we should have used $\ln(-P(t))$. But one can easily prove that the formula $P(t) = P(0)e^{kt}$ is always valid. In any case, for population problems, P is positive.)

Which is better in practice, to use the difference equation (A.2) or the differential equation (A.3)? It is hard to say: the answer depends on the application. Mathematically, differential equations are usually easier to analyze, although sometimes, as when we study chaotic behavior in simple one-dimensional systems, difference equations may give great insight. Also, we often use difference equations as a basis of numerical techniques which allow us to find an approximation of the solution of a differential equation. For example, Euler's method basically reverses the process of going from (A.1) to (A.3).

Let us now look at some more examples of differential equations.

Limits to Growth: Logistic Equation

Often, there are limits imposed by the environment on the maximal possible size of a population: not enough nutrients for a large bacterial culture, insufficient food for the human population of an island, or a small hunting territory for a given animal species. Ecologists talk about the *carrying capacity* of the environment, a number N with the property that no populations $P > N$ are sustainable. If the population starts bigger than N, the number of individuals will decrease. To come up with an equation that represents this situation, we follow the same steps that we did before, except that now we have that $P(t + \Delta t) - P(t)$ should be negative if $P(t) > N$. In other words, we have $P(t+\Delta t)-P(t) = f(P(t))\Delta t$, where $f(P)$ is not just "kP" but should be instead a more complicated expression involving P, and which has the properties that:

- $f(0) = 0$ (no increase in the population if there is no one around to start with!),

- $f(P) > 0$ when $0 < P < N$ (the population increases while there are enough resources), and

- $f(P) < 0$ when $P > N$.

Taking limits just like we did before, we arrive to the differential equation:

$$P'(t) = f(P(t)).$$

From now on, we will drop the "t" when it is obvious, and use the shorthand notation $P' = f(P)$ instead of the more messy $P'(t) = f(P(t))$. We must still decide what function "f" is appropriate. Because of the properties wanted ($f(0) = 0$, $f(P) > 0$ when $0 < P < N$, $f(P) < 0$ when $P > N$), the simplest choice is a parabola which opens downward and has zeroes at $P = 0$ and $P = N$: $f(P) = -cP(P - N)$, with $c > 0$, or, with $k = cN$, $f(P) = kP(1 - P/N)$. We arrive in this way to the *logistic population model*

$$P' = kP \left(1 - \frac{P}{N} \right). \tag{A.4}$$

(Remember: this is shorthand for $P'(t) = kP(t)(1 - P(t)/N)$.) The constant k is positive, since it was obtained as cN.

Solution of Logistic Equation

Like $P' = kP$, equation (A.4) is one of those (comparatively few) equations which can actually be solved in closed form. To solve it, we do almost the same that we did with $P' = kP$ (this is an example of the method of *separation of variables*): we write the equation as $dP/dt = kP(1 - (P/N))$,

formally multiply both sides by dt and divide by $P(1 - (P/N))$, arriving at

$$\frac{dP}{P(1 - P/N)} = k.$$

Next we take antiderivatives of both sides, obtaining

$$\int \frac{dP}{P(1 - P/N)} = \int kdt.$$

The right-hand side can be evaluated using partial fractions:

$$\frac{1}{P(1 - P/N)} = \frac{N}{P(N - P)} = \frac{1}{P} + \frac{1}{N - P}$$

so

$$\ln P - \ln(N - P) + c_1 = kt + c_2$$

for some constants c_1 and c_2, or, with $c = c_2 - c_1$,

$$\ln \left(\frac{P}{N - P} \right) = kt + c \tag{A.5}$$

and, taking exponentials of both sides,

$$\frac{P}{N - P} = Ce^{kt} \tag{A.6}$$

with $C = e^c$. This is an algebraic equation for P, but we can go a little further and solve explicitly:

$$P = Ce^{kt}(N - P) \Rightarrow Ce^{kt}P + P = Ce^{kt}N \Rightarrow P = \frac{Ce^{kt}N}{Ce^{kt} + 1} = \frac{N}{1 + \frac{1}{C}e^{-kt}}.$$

Finally, to find C, we can evaluate both sides of equation (A.6) at $t = 0$:

$$C = \frac{P(0)}{N - P(0)}$$

and therefore conclude that

$$P(t) = \frac{P(0)N}{P(0) + (N - P(0))e^{-kt}}. \tag{A.7}$$

Observe that, since $e^{-kt} \to 0$ as $t \to \infty$, $P(t) \to N$, which is not surprising. (Why?)

This formula is also valid for negative values of t with $P(t) \to 0$ as $t \to -\infty$.

Homework assignment: use a computer to plot several solutions of the equation, for various values of N and of $P(0)$.

Some "Small-Print Legal Disclaimers"

(You may want to skip this section in a first reading.)

We cheated a bit when deriving the solution for the logistic equation. First of all, we went a bit too fast over the "divide by dt" business. What is the meaning of dividing by the differential? Well, it turns out that it is OK to do this, because what we did can be interpreted as, basically, just a way of applying (backwards) the chain rule. Let us justify the above steps without using differentials. Starting from the differential equation (A.4) we can write, assuming that $P \neq 0$ and $P \neq N$ (so that we are not dividing by zero):

$$\frac{P'}{P(1 - P/N)} = k. \tag{A.8}$$

Now, one antiderivative of $1 \,/\, (P(1 - P/N))$, as a function of P, is the function

$$Q(P) = \ln \left(P \,/\, (N - P) \right)$$

(let us suppose that $N > P$, so the expression inside the log is positive). So, the chain rule says that

$$\frac{dQ(P(t))}{dt} = \frac{dQ}{dP} \frac{dP}{dt} = \frac{1}{P(1 - P/N)} P'(t).$$

Therefore, equation (A.8) gives us that

$$\frac{dQ(P(t))}{dt} = k$$

from which we then conclude, by taking antiderivatives, that

$$Q(P(t)) = kt + c$$

which is exactly the same as the equation (A.5) which had before been obtained using differentials. In general, we can always justify "separation of variables" solutions in this manner, but from now on we will skip this step and use the formal method.

There is still a small gap in our arguments, namely we assumed that $P \neq 0$ and that $P \neq N$ (so that we were not dividing by zero) and also $N > P$, so the expression inside the log was positive.

There is a theorem that states that, under appropriate conditions (differentiability of f), solutions are unique. Thus, since $P = 0$ and $P = N$ are equilibria, any solution that starts with $P(0) > N$ will always have $P(t) > N$, and a similar property is true for each of the intervals $P < 0$ and $0 < P < N$. So we can treat each of the cases separately.

If $N < P$, then the antiderivative is $\ln |P \,/\, (N - P)|$ (that is, we use absolute values). But this doesn't change the general solution. All it means is that equation (A.6) becomes

$$\left| \frac{P}{N - P} \right| = Ce^{kt}$$

which can also be written as in (A.6) but with C negative. We can treat the case $P < 0$ in the same way.

Finally, the exceptional cases when P could be zero or N are taken care of once we notice that the general solution (A.7) makes sense when $P(0) = 0$ (we get $P \equiv 0$) or when $P(0) = N$ (we get $P \equiv N$).
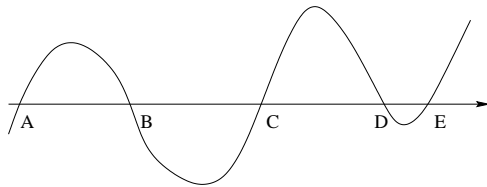
Equilibria

Observe that if, for some time t_0, it happens that $P(t_0) = 0$, then the right-hand side of the differential equation (A.4) becomes zero, so $P'(t_0) = 0$, which means that the solution cannot "move" from that point. So the value $P = 0$ is an **equilibrium point** for the equation: a value with the property that if we start there, then we stay there forever. This is not a particularly deep conclusion: if we start with zero population we stay with zero population. Another root of the right hand side is $P = N$. If $P(t_0) = N$ then $P'(t_0) = 0$, so if we start with exactly N individuals, the population also remains constant, this time at N. Again, this is not surprising, since the model was derived under the assumption that populations larger than N decrease and populations less than N increase.

In general, for any differential equation of the form $y' = f(y)$, we say that a point $y = a$ is an *equilibrium* if a is a root of f, that is, $f(a) = 0$. This means that if we start at $y = a$, we cannot move away from $y = a$. Or, put in a different way, the constant function $y(t) \equiv a$ is a solution of $y' = f(y)$ (because $y'(t) = a' \equiv 0$ and also $f(y(t)) = f(a) = 0$. One says also that the constant function $y(t) = a$ is an *equilibrium solution* of $y' = f(y)$.

The analysis of equilibria allows us to obtain a substantial amount of information about the solutions of a differential equation of the type $y' = f(y)$ with very little effort, in fact without even having to solve the equation. (For "nonautonomous" equations, when t appears in the right hand side: $y' = f(t, y)$, this method doesn't quite work, because we need to plot f against two variables. The technique of slope fields is useful in that case.) The fundamental fact that we need is that — assuming that f is a differentiable function — *no trajectory can pass through an equilibrium*: if we are ever at an equilibrium, we must have always been there and we will remain there forever. This will be explained later, when covering uniqueness of solutions.

For example, suppose that we know that the plot of $f(y)$ against y looks like this:

where we labeled the points where $f(y)$ has roots, that is to say, the equilibria of $y' = f(y)$.

We can conclude that any solution $y(t)$ of $y' = f(P)$ which starts just to the right of A will move rightwards, because $f(y)$ is positive for all points between A and B, and so $y' > 0$. Moreover, we cannot cross the equilibrium B, so any such trajectory stays in the interval (A, B) and, as t increases, it approaches asymptotically the point B. To summarize, if $y(0) = y_0$ with $y_0 \in (A, B)$, then the graph of
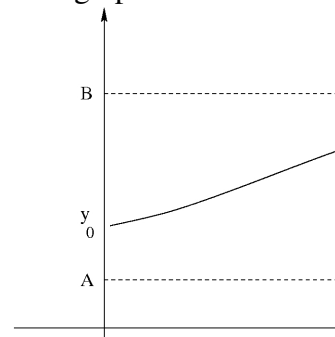
the solution $y(t)$ of $y' = f(y)$ must look more or less like this:

Homework assignment: For the same function f shown above, give an approximate plot of a solution

of $y' = f(y)$ for which $y(0) \in (B, C)$. Repeat with $y(0) \in (C, D)$ and with $y(0) \in (D, E)$.

Systems

More generally, one considers *systems* of differential equations, such as for example:

$$\frac{dx}{dt} = 2\,x - 5\,xy$$

$$\frac{dy}{dt} = -y + 1.2\,xy\,.$$

This example might represent the number of individuals of each of two species of animals, in which the "y" species is a predator of "x". The first species reproduces (at rate "2") if there are no y's present, but when there are y's around, there is a "death rate" for x that is proportional to the number of predators. Similarly, the second population grows in proportion to the population size of x's, but it diminishes when there are no x's (its only source of nutrition).

More Examples

Let us discuss some more easy examples (of single-variable problems).

(a) Populations under Harvesting

Let us return to the population model (A.4):

$$P' = kP\left(1 - \frac{P}{N}\right)$$

which describes population growth under environmental constraints. Suppose that $P(t)$ represents the population of a species of fish, and that fishing removes a certain number K of fish each unit of time. This means that there will be a term in $P(t + \Delta t) - P(t)$ equal to $-K\Delta t$. When we divide by Δt and take limits, we arrive at the equation for resources under constant harvesting:

$$P' = kP\left(1 - \frac{P}{N}\right) - K.$$

Many variations are possible. For example, it is more realistic to suppose that a certain proportion of fish are caught per unit of time (the more fish, the easier to catch). This means that, instead of a term $-K\Delta t$ for how many fish are taken away in an interval of length Δt, we'd now have a term of the form $-KP(t)\Delta t$, which is proportional to the population. The differential equation that we obtain is now $P' = kP(1 - (P/N)) - KP$. Or, if only fish near the surface can be caught, the proportion of fish caught per unit of time may depend on the power $P^{2/3}$ (do you understand why? are you sure?). This would give us the equation $P' = kP(1 - (P/N)) - KP^{2/3}$.

(b) Epidemics

The spread of epidemics is another example whose study can be carried out using differential equations. Suppose that $S(t)$ counts the number of individuals infected with a certain virus, at time t, and that people mix randomly and get infected from each other if they happen to be close. One model is as follows. The increase in the number of infected individuals $S(t + \Delta t) - S(t)$ during a time interval of length Δt is proportional to the number of close encounters between sick and healthy individuals, that is, to $S(t)H(t)\Delta t$, because $S(t)H(t)$ is the total number of pairs of (sick,healthy) individuals, and the longer the interval, the more chances of meeting. Taking limits as usual, we arrive to $S'(t) = kS(t)H(t)$, where k is some constant. If the total number of individuals is N, then $H(t) = N - S(t)$, and the equation becomes:

$$S' = kS(t)(N - S(t))$$

which is a variant of the logistic equation. There are many extensions of this idea. For instance, if in every Δt time interval a certain proportion of infected individuals get cured, we'd have a term $-kS(t)$.

(c) Chemical Reactions

Chemical reactions also give rise to similar models. Let us say that there are two reactants A and B, which may combine to give C via $A + B \to C$ (for each molecule of A and B, we obtain a molecule of C). If the chemicals are well-mixed, the chance of two molecules combining is proportional to how many pairs there are and to the length of time elapsed (just like with the infection model, molecules need to get close enough to react). So $c'(t) = ka(t)b(t)$, where $a(t)$ is the amount of A at time t and $b(t)$ the amount of B. If we start with amounts a_0 and b_0 respectively, and we have $c(t)$ molecules of C at time t, this means that $a(t) = a_0 - c(t)$ and $b(t) = b_0 - c(t)$, since one molecule of A and B was used up for each molecule of C that was produced. So the equation becomes

$$c' = k(a_0 - c)(b_0 - c).$$

(d) Air Resistance

Consider a body moving in air (or another fluid). For low speeds, air resistance (drag) is proportional to the speed of the object, and acts to slow down the object, in other words, it acts as a force $k\,|v|$, in a direction opposite to movement, where $|v|$ is the absolute value of the velocity. Suppose that a body is falling towards the earth, and let us take "down" as the positive direction of movement. In that case, Newton's "$F = ma$" law says that the mass times the acceleration v' is equal to the total force on the body, namely mg (its weight) plus the effect of drag, which is $-kv$ (because the force acts opposite to the direction of movement):

$$mv' = mg - kv.$$

For large velocities, drag is often modeled more accurately by a quadratic effect $-kv^2$ in a direction opposite to movement. This would lead to an equation like $mv' = mg - kv^2$ for the velocity of a falling object. Both of these equations can be solved exactly. This allows the validity of the model to be tested by comparing these formulas to experimental results.

(e) Newton's Law of Cooling

The temperature inside a building is assumed to be uniform (same in every room) and is given by $y(t)$ as a function of the time t. The outside air is at temperature $a(t)$, which also depends on the time of the day, and there is a furnace which supplies heat at a rate $h(t)$ (or, for negative h, an air-conditioning unit which removes heat at that rate). What is the temperature in the building? Newton's law of cooling tells us that the rate of change of temperature dy/dt will depend on the difference between the inside and outside temperatures (the greater the difference, the faster the change), with a term added to model the effect of the furnace:

$$mcy' = -k(y - a(t)) + h(t),$$

where the mass of air in the building is the constant m (no windows can be opened, and doors are usually tightly closed, being opened rarely and briefly, so we assume that m is a constant), c is a positive constant (the heat capacity), and k is another positive constant (which is determined by insulation, building layout, etc).

Homework Problem

You should match the following word descriptions and differential equations. *More than one equation may match a description, and vice versa.*

Descriptions:

1. The rate of change of the population of a certain country, which depends on the birth and death rates as well as on the number of immigrants, who arrive at a constant rate into the country.

2. The rate of change of the population of a certain country, which depends on the birth and death rates, but there is a net emigration from the country (at a constant rate).

3. Fish in a certain area, which reproduce in proportion to the population, subject to limits imposed by the carrying capacity of the environment, and the population of which is also reduced by fishing which proceeds at a constant rate.

4. The temperature of a building, when the outside temperature varies periodically (it goes down during the night, up during the day) and there is no heating or air-conditioning.

5. The temperature of a building, when the outside temperature varies periodically (it goes down during the night, up during the day) and heating is being applied at a constant rate.

6. The temperature of a building, when the outside temperature is constant, and there is no heating or air-conditioning.

7. The temperature of a building, when the outside temperature is constant, and heating is being applied at a constant rate.

8. The amount of money in a savings account, when interest is compounded continuously, and also additional money is being added at a constant rate (the person always deposits a certain percentage of her paycheck).

9. The rate of change of the volume of a raindrop, which evaporates at a rate proportional to its surface area.

10. The rate of change of the volume of a raindrop, which evaporates at a rate proportional to its diameter.

11. The mass of a radioactive substance which is decaying (at a rate proportional to the amount present).

12. The amount of chlorine in a swimming pool; chlorinated water is added at a fixed rate, the water in the pool is well-mixed, and water is being removed from the pool so that the total volume is constant.

Equations (all constants are positive):

- $y' = -ky$ Answer(s):_____

- $y' = -ky + c$ Answer(s):_____

- $y' = -ky^{1/3}$ Answer(s):_____

- $y' = -ky^{2/3}$ Answer(s):_____

- $y' = ky(K - y)$ Answer(s):_____

- $y' = ky(K - y) + c$ Answer(s):_____

- $y' = ky(K - y) - c$ Answer(s):_____

- $y' = -k(y - \sin t) + c$ Answer(s):_____

- $y' = -k(y - \sin t)$ Answer(s):_____

- $y' = -k(y - \sin t) - c$ Answer(s):_____

- $y' = -k(y - K) + c$ Answer(s):_____

- $y' = -k(y - K) - c$ Answer(s):_____

- $y' = -k(y - K)$ Answer(s):_____

- $y' = ky$ Answer(s):_____

- $y' = ky + c$ Answer(s):_____

- $y' = ky - c$ Answer(s):_____

A.2 Phase-planes

A technique which is often very useful in order to analyze the phase plane behavior of a two-dimensional autonomous system

$$
\begin{aligned}
\frac{dx}{dt} &= f(x, y) \\
\frac{dy}{dt} &= g(x, y)
\end{aligned}
$$

is to attempt to understand the graphs of solutions $(x(t), y(t))$ as the level sets of some function $h(x, y)$.

Some Examples.

Example 1

For example, take

$$
\begin{aligned}
\frac{dx}{dt} &= -y \\
\frac{dy}{dt} &= x
\end{aligned}
$$

(that is, $f(x, y) = -y$ and $g(x, y) = x$). If we could solve for t as a function of x, by inverting the function $x(t)$, and substitute the expression that we obtain into $y(t)$, we would end up with an expression $y(x)$ for the y-coordinate in terms of the x coordinate, eliminating t. This cannot be done in general, but it suggests that we may want to look at dy/dx. Formally (or, more precisely, using the chain rule), we have that

$$
\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{x}{-y}
$$

which is a differential equation for y as a variable dependent on x. This equation is separable:

$$
\int \frac{dy}{y} = \int -\frac{dx}{x}
$$

so we obtain, taking antiderivatives,

$$
\frac{y^2}{2} = -\frac{x^2}{2} + c
$$

where c is an undetermined constant, and since c must be nonnegative, we can write $c = r^2$. *In conclusion, the solutions $(x(t), y(t))$ all lie in the circles $x^2 + y^2 = r^2$ of different radii and centered at zero.* Observe that we have **not** solved the differential equation, since we did determine the forms of x and y as functions of t (which, as a matter of fact, are trigonometric functions $x = r \cos t$, $y = r \sin t$ that draw circles at constant unit speed in the counterclockwise direction). What we have done is just to find curves (the above-mentioned circles) which contain all solutions. Even though this is less interesting (perhaps) than the actual solutions, it is still very interesting. We know what the general phase plane picture looks like.

A variation of this example is:

$$\frac{dx}{dt} = -y$$
$$\frac{dy}{dt} = 2x$$

for which it is easy to see, by a similar reasoning, that the solutions line in ellipses of the form

$$x^2 + \frac{y^2}{2} = r^2 .$$

Example 2

Another example is this:

$$\frac{dx}{dt} = y^5 e^x$$
$$\frac{dy}{dt} = x^5 e^x.$$

Here, $dy/dx = x^5/y^5$ so we get again a separable equation, and we see that the solutions all stay in the curves

$$x^6 - y^6 = c.$$

Example 3

More interesting is the general case of predator-prey equations:

$$\frac{dx}{dt} = ax - bxy$$
$$\frac{dy}{dt} = -cy + dxy$$

where a, b, c, d are all positive constants. Then

$$\frac{dy}{dx} = \frac{y(-c + dx)}{x(a - by)}$$

so

$$\int \left(\frac{a}{y} - b\right) dy = \int \left(-\frac{c}{x} + d\right) dx$$

and from here we conclude that the solutions all stay in the sets

$$a \ln(y) - by + c \ln(x) - dx = k$$

for various values of the constant k. It is not obvious what these sets look like, but if you graph the level sets of the function

$$h(x, y) = a \ln(y) - by + c \ln(x) - dx$$

you'll see that the level sets look like the orbits of the predator-prey system shown, for the special values $a = 2$, $b = 1.2$, $c = 1$, and $d = 0.9$ shown in the *Maple* plot below
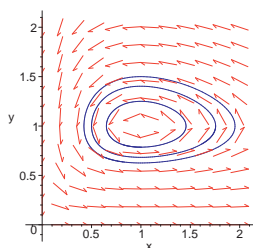
using the values $a = b = 6$ and $c = d = 2$. The initial values for these three curves were $(1, 1.25)$, $(0.5, 1)$, and $(1, 1, 5)$.

(Of course, the scales will be different for different values of the constants, but the picture will look the same, in general terms.) This argument is used to prove that predator-prey systems always lead to periodic orbits, no matter what the coefficients of the equation are.

Homework

In each of the following problems, a system

$$\frac{dx}{dt} = f(x, y)$$
$$\frac{dy}{dt} = g(x, y)$$

is given. Solve the equation

$$\frac{dy}{dx} = \frac{g(x, y)}{f(x, y)}$$

and use the information to sketch what the orbits of the original equation should look like. Exercise 1

$$\frac{dx}{dt} = y(1 + x^2 + y^2)$$
$$\frac{dy}{dt} = x(1 + x^2 + y^2)$$

Exercise 2

$$\frac{dx}{dt} = 4y(1 + x^2 + y^2)$$
$$\frac{dy}{dt} = \frac{dy}{dt} = -x(1 + x^2 + y^2)$$

Exercise 3

$$\frac{dx}{dt} = y^3 e^{x+y}$$
$$\frac{dy}{dt} = -x^3 e^{x+y}$$

Exercise 4

$$\frac{dx}{dt} = y^2$$
$$\frac{dy}{dt} = (2x+1)y^2$$

Exercise 5

$$\frac{dx}{dt} = e^{xy}\cos(x)$$
$$\frac{dy}{dt} = e^{xy}$$

A.3   Matrix Exponentials

Generalities

A system of autonomous linear differential equations can be written as

$$\frac{dY}{dt} = AY$$

where A is an n by n matrix and $Y = Y(t)$ is a vector listing the n dependent variables. (In most of what we'll do, we take $n = 2$, since we study mainly systems of 2 equations, but the theory is the same for all n.)

If we were dealing with just one linear equation

$$y' = ay$$

then the general solution of the equation would be e^{at}. It turns out that *also for vector equations the solution looks like this, provided that we interpret what we mean by "e^{At}" when A is a matrix instead of just a scalar.* How to define e^{At}? The most obvious procedure is to take the power series which defines the exponential, which as you surely remember from Calculus is

$$e^x = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \cdots + \frac{1}{k!}x^k + \cdots$$

and just formally plug-in $x = At$. (The answer should be a matrix, so we have to think of the term "1" as the identity matrix.) In summary, we *define*:

$$e^{At} = I + At + \frac{1}{2}(At)^2 + \frac{1}{6}(At)^3 + \cdots + \frac{1}{k!}(At)^k + \cdots$$

where we understand the series as defining a series for each coefficient. One may prove that:

$$e^{A(t+s)} = e^{At}e^{As} \quad \text{for all} \ \ s, t. \tag{A.9}$$

and therefore, since (obviously) $e^{A0} = I$, using $s = -t$ gives

$$e^{-At} = \left(e^{At}\right)^{-1} \tag{A.10}$$

(which is the matrix version of $e^{-x} = 1/e^x$). We now prove that this matrix exponential has the following property:

$$\frac{de^{At}}{dt} = Ae^{At} = e^{At}A \tag{A.11}$$

for every t.

Proof Let us differentiate the series term by term:

$$\begin{aligned}
\frac{de^{At}}{dt} &= \frac{d}{dt}\left(I + At + \frac{1}{2}(At)^2 + \frac{1}{6}(At)^3 + \cdots + \frac{1}{k!}(At)^k + \cdots\right) \\
&= 0 + A + A^2 t + \frac{1}{2}A^3 t^2 + \cdots + \frac{1}{(k-1)!}A^k t^{k-1} + \cdots \\
&= A\left(I + At + \frac{1}{2}(At)^2 + \frac{1}{6}(At)^3 + \cdots + \frac{1}{k!}(At)^k + \cdots\right) \\
&= Ae^{At}
\end{aligned}$$

and a similar proof, factoring A on the right instead of to the left, gives the equality between the derivative and $e^{At}A$. (Small print: the differentiation term-by-term can be justified using facts about term by term differentiation of power series inside their domain of convergence.) The property (A.11) is the fundamental property of exponentials of matrices. It provides us immediately with this corollary:

The initial value problem $\dfrac{dY}{dt} = AY$, $Y(0) = Y_0$ *has the unique solution* $Y(t) = e^{At}Y_0$.

We can, indeed, verify that the formula $Y(t) = e^{At}Y_0$ defines a solution of the IVP:

$$\frac{dY(t)}{dt} = \frac{de^{At}Y_0}{dt} = \frac{de^{At}}{dt}Y_0 = \left(Ae^{At}\right)Y_0 = A\left(e^{At}Y_0\right) = AY(t).$$

(That it is the unique, i.e., the only, solution is proved as follows: if there were another solution $Z(t)$ of the same IVP, then we could let $W(t) = Y(t) - Z(t)$ and notice that $W' = Y' - Z' = A(Y - Z) = AW$, and $W(0) = Y(0) - Z(0) = 0$. Letting $V(t) = e^{-At}W(t)$, and applying the product rule, we have that

$$V' = -Ae^{-At}W + e^{-At}W' = -e^{-At}AW + e^{-At}AW = 0$$

so that V must be constant. Since $V(0) = W(0) = 0$, we have that V must be identically zero. Therefore $W(t) = e^{At}V(t)$ is also identically zero, which because $W = Y - Z$, means that the functions Y and Z are one and the same, which is what we claimed.)

Although we started by declaring Y to be a vector, the equation $Y' = AY$ makes sense as long as Y can be multiplied on the left by A, i.e., whenever Y is a matrix with n rows (and any number of columns). In particular, e^{At} itself satisfies this equation. The result giving uniqueness of solutions of initial value problems applies to matrices since each column satisfies the equation and has the corresponding column of the initial data as its initial value. The value of e^{At} at $t = 0$ is the n by n identity matrix. This initial value problem characterizes e^{At}. Verification of these properties is an excellent check of a calculation of e^{At}.

So we have, in theory, solved the general linear differential equation. A potential problem is, however, that it is not always easy to calculate e^{At}.

Some Examples

We start with this example:

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}. \tag{A.12}$$

We calculate the series by just multiplying A by t:

$$At = \begin{pmatrix} t & 0 \\ 0 & 2t \end{pmatrix}$$

and now calculating the powers of At. Notice that, because At is a diagonal matrix, its powers are very easy to compute: we just take the powers of the diagonal entries *(why? if you don't understand,* **stop** *and think it over right now)*. So, we get

$$e^{At} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} t & 0 \\ 0 & 2t \end{pmatrix} + \frac{1}{2}\begin{pmatrix} t^2 & 0 \\ 0 & (2t)^2 \end{pmatrix} + \frac{1}{6}\begin{pmatrix} t^3 & 0 \\ 0 & (2t)^3 \end{pmatrix} + \cdots + \frac{1}{k!}\begin{pmatrix} t^k & 0 \\ 0 & (2t)^k \end{pmatrix} + \cdots$$

and, just adding coordinate-wise, we obtain:

$$e^{At} = \begin{pmatrix} 1 + t + \frac{1}{2}t^2 + \frac{1}{6}t^3 + \cdots + \frac{1}{k!}t^k + \cdots & 0 \\ 0 & 1 + 2t + \frac{1}{2}(2t)^2 + \frac{1}{6}(2t)^3 + \cdots + \frac{1}{k!}(2t)^k + \cdots \end{pmatrix}$$

which gives us, finally, the conclusion that

$$e^{At} = \begin{pmatrix} e^t & 0 \\ 0 & e^{2t} \end{pmatrix}.$$

So, in this very special case we obtained the exponential by just taking the exponentials of the diagonal elements and leaving the off-diagonal elements zero (observe that we did not end up with exponentials of the non-diagonal entries, since $e^0 = 1$, not 0).

In general, computing an exponential is a little more difficult than this, and it is not enough to just take exponentials of coefficients. Sometimes things that seem surprising (the first time that you see them) may happen. Let us take this example now:

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{A.13}$$

To start the calculation of the series, we multiply A by t:

$$At = \begin{pmatrix} 0 & t \\ -t & 0 \end{pmatrix}$$

and again calculate the powers of At. This is a little harder than in the first example, but not too hard:

$$(At)^2 = \begin{pmatrix} -t^2 & 0 \\ 0 & -t^2 \end{pmatrix}$$

$$(At)^3 = \begin{pmatrix} 0 & -t^3 \\ t^3 & 0 \end{pmatrix}$$

$$(At)^4 = \begin{pmatrix} t^4 & 0 \\ 0 & t^4 \end{pmatrix}$$

$$(At)^5 = \begin{pmatrix} 0 & t^5 \\ -t^5 & 0 \end{pmatrix}$$

$$(At)^6 = \begin{pmatrix} -t^6 & 0 \\ 0 & -t^6 \end{pmatrix}$$

and so on. We won't compute more, because by now you surely have recognized the pattern (*right?*). We add these up (not forgetting the factorials, of course):

$$e^{At} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & t \\ -t & 0 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} -t^2 & 0 \\ 0 & -t^2 \end{pmatrix} + \frac{1}{3!}\begin{pmatrix} 0 & -t^3 \\ t^3 & 0 \end{pmatrix} + \frac{1}{4!}\begin{pmatrix} t^4 & 0 \\ 0 & t^4 \end{pmatrix} + \cdots$$

and, just adding each coordinate, we obtain:

$$e^{At} = \begin{pmatrix} 1 - \frac{t^2}{2} + \frac{t^4}{4!} - \cdots & t - \frac{t^3}{3!} + \frac{t^5}{5!} - \cdots \\ -t + \frac{t^3}{3!} - \frac{t^5}{5!} + \cdots & 1 - \frac{t^2}{2} + \frac{t^4}{4!} - \cdots \end{pmatrix}$$

which gives us, finally, the conclusion that

$$e^{\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} t} = e^{At} = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}.$$

It is remarkable that trigonometric functions have appeared. Perhaps we made a mistake? How could we make sure? Well, let us *check* that property (A.11) holds (we'll check only the first equality, you can check the second one). We need to test that

$$\frac{d}{dt} \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} = A \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}. \qquad (A.14)$$

Since

$$\frac{d}{dt}(\sin t) = \cos t, \quad \text{and} \quad \frac{d}{dt}(\cos t) = -\sin t,$$

we know that

$$\frac{d}{dt} \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} = \begin{pmatrix} -\sin t & \cos t \\ -\cos t & -\sin t \end{pmatrix}$$

and, on the other hand, multiplying matrices:

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} = \begin{pmatrix} -\sin t & \cos t \\ -\cos t & -\sin t \end{pmatrix}$$

so we have verified the equality (A.14).

As a last example, let us take this matrix:

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}. \qquad (A.15)$$

Again we start by writing

$$At = \begin{pmatrix} t & t \\ 0 & t \end{pmatrix}$$

and calculating the powers of At. It is easy to see that the powers are:

$$(At)^k = \begin{pmatrix} t^k & kt^k \\ 0 & t^k \end{pmatrix}$$

since this is obviously true for $k = 1$ and, recursively, we have

$$(At)^{k+1} = (At)^k A = \begin{pmatrix} t^k & kt^k \\ 0 & t^k \end{pmatrix} \begin{pmatrix} t & t \\ 0 & t \end{pmatrix} = \begin{pmatrix} t^k t & t^k t + kt^k t \\ 0 & t^k t \end{pmatrix} = \begin{pmatrix} t^{k+1} & (k+1)t^{k+1} \\ 0 & t^{k+1} \end{pmatrix}.$$

Therefore,

$$\begin{aligned} e^{At} &= \sum_{k=0}^{\infty} \begin{pmatrix} t^k/k! & kt^k/k! \\ 0 & t^k/k! \end{pmatrix} \\ &= \begin{pmatrix} \sum_{k=0}^{\infty} \dfrac{t^k}{k!} & \sum_{k=0}^{\infty} \dfrac{kt^k}{k!} \\ 0 & \sum_{k=0}^{\infty} \dfrac{t^k}{k!} \end{pmatrix} \\ &= \begin{pmatrix} e^t & te^t \\ 0 & e^t \end{pmatrix}. \end{aligned}$$

To summarize, we have worked out three examples:

- The first example (A.12) is a diagonal matrix, and we found that its exponential is obtained by taking exponentials of the diagonal entries.

- The second example (A.13) gave us an exponential matrix that was expressed in terms of trigonometric functions. Notice that this matrix has imaginary eigenvalues equal to i and $-i$, where $i = \sqrt{-1}$.

- The last example (A.15) gave us an exponential matrix which had a nonzero function in the $(1, 2)$-position. Notice that this nonzero function was *not* just the exponential of the $(1, 2)$-position in the original matrix. That exponential would give us an e^t term. Instead, we got a more complicated te^t term.

In a sense, these are all the possibilities. Exponentials of all two by two matrices can be obtained using functions of the form e^{at}, te^{at}, and trigonometric functions (possibly multiplied by e^{at}). Indeed, exponentials of any size matrices, not just 2 by 2, can be expressed using just polynomial combinations of t, scalar exponentials, and trigonometric functions. We will not quite prove this fact here; you should be able to find the details in any linear algebra book.

Calculating exponentials using power series is OK for very simple examples, and important to do a few times, so that you understand what this all means. But in practice, one uses very different methods for computing matrix exponentials. (Remember how you first saw the definition of derivative using limits of incremental quotients, and computed some derivatives in this way, but soon learned how to use "the Calculus" to calculate derivatives of complicated expressions using the multiplication rule, chain rule, and so on.)

Computing Matrix Exponentials

We wish to calculate e^{At}. The key concept for simplifying the computation of matrix exponentials is that of *matrix similarity*. Suppose that we have found two matrices, Λ and S, where S is invertible, such that this formula holds:

$$A = S\Lambda S^{-1} \tag{A.16}$$

(if (A.16) holds, one says that A and Λ are similar matrices). Then, we claim, it is true that also:

$$e^{At} = S\, e^{\Lambda t}\, S^{-1} \tag{A.17}$$

for all t. Therefore, if the matrix Λ is one for which $e^{\Lambda t}$ is easy to find (for example, if it is a diagonal matrix), we can then multiply by S and S^{-1} to get e^{At}. To see why (A.17) is a consequence of (A.16), we just notice that $At = S(\Lambda t)S^{-1}$ and we have the following "telescopic" property for powers:

$$(At)^k = \big(S(\Lambda t)S^{-1}\big)\big(S(\Lambda t)S^{-1}\big)\cdots\big(S(\Lambda t)S^{-1}\big) = S(\Lambda t)^k S^{-1}$$

since the terms may be regrouped so that all the in-between pairs $S^{-1}S$ cancel out. Therefore,

$$
\begin{aligned}
e^{At} &= I + At + \frac{1}{2}(At)^2 + \frac{1}{6}(At)^3 + \cdots + \frac{1}{k!}(At)^k + \cdots \\
&= I + S(\Lambda t)S^{-1} + \frac{1}{2}S(\Lambda t)^2 S^{-1} + \frac{1}{6}S(\Lambda t)^3 S^{-1} + \cdots + \frac{1}{k!}S(\Lambda t)^k S^{-1} + \cdots \\
&= S\left[I + \Lambda t + \frac{1}{2}(\Lambda t)^2 + \frac{1}{6}(\Lambda t)^3 + \cdots + \frac{1}{k!}(\Lambda t)^k + \cdots\right]S^{-1} \\
&= S e^{\Lambda t} S^{-1}
\end{aligned}
$$

as we claimed.

The basic theorem is this one:

Theorem. For every n by n matrix A with entries in the complex numbers, one can find an invertible matrix S, and an upper triangular matrix Λ such that (A.16) holds.

Remember that an upper triangular matrix is one that has the following form:

$$\begin{pmatrix} \lambda_1 & * & * & \cdots & * & * \\ 0 & \lambda_2 & * & \cdots & * & * \\ 0 & 0 & \lambda_2 & \cdots & * & * \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_{n-1} & * \\ 0 & 0 & 0 & \cdots & 0 & \lambda_n \end{pmatrix}$$

where the stars are any numbers. The numbers $\lambda_1, \ldots, \lambda_n$ turn out to be the eigenvalues of A.

There are two reasons that this theorem is interesting. First, it provides a way to compute exponentials, because it is not difficult to find exponentials of upper triangular matrices (the example (A.15) is actually quite typical) and second because it has important theoretical consequences.

Although we don't need more than the theorem stated above, there are two stronger theorems that you may meet elsewhere. One is the "Jordan canonical form" theorem, which provides a matrix Λ that is not only upper triangular but which has an even more special structure. Jordan canonical forms are theoretically important because they are essentially unique (that is what "canonical" means in this context). Hence, the Jordan form allows you to determine whether or not two matrices are similar. However, it is not very useful from a computational point of view, because they are what is known in numerical analysis as "numerically unstable", meaning that small perturbations of A can give one totally different Jordan forms. A second strengthening is the "Schur unitary triangularization theorem" which says that one can pick the matrix S to be *unitary*. (A unitary matrix is a matrix with entries in the complex numbers whose inverse is the complex conjugate of its transpose. For matrices S with real entries, then we recognize it as an *orthogonal* matrix. For matrices with complex entries, unitary matrices turn out to be more useful than other generalization of orthogonal matrices that one may propose.) Schur's theorem is extremely useful in practice, and is implemented in many numerical algorithms.

We do not prove the theorem here in general, but only show it for $n = 2$; the general case can be proved in much the same way, by means of a recursive process.

We start the proof by remembering that every matrix has at least one eigenvalue, let us call it λ, and an associate eigenvector, v. That is to say, v is a vector **different from zero**, and

$$Av = \lambda v . \tag{A.18}$$

If you stumble on a number λ and a vector v that you believe to an eigenvalue and its eigenvector, you should *immediately* see if (A.18) is satisfied, since that is an easy calculation. Numerical methods for finding eigenvalues and eigenvectors take this approach.

For theoretical purposes, it is useful to note that the the eigenvalues λ can be characterized as the roots of the characteristic equation

$$\det(\lambda I - A) = 0.$$

For two-dimensional systems, this is the same as the equation

$$\lambda^2 - \text{trace}\,(A)\lambda + \det(A) = 0$$

with

$$\text{trace} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = a + d$$

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc.$$

Now, quadratic equations are easy to solve, so this approach is also computationally useful for 2 by 2 matrices.

There are, for 2 by 2 matrices with *real* entries, either two real eigenvalues, one real eigenvalue with multiplicity two, or two complex eigenvalues. In the last case, the two complex eigenvalues must be conjugates of each other.

If you have λ, an eigenvector associated to an eigenvalue λ is then found by solving the linear system

$$(A - \lambda I)v = 0$$

(*since λ is a root of the characteristic equation*, there are an infinite number of solutions; we pick any nonzero one).

With an eigenvalue λ and eigenvector v found, we next pick *any* vector w with the property that the two vectors v and w are linearly independent. For example, if

$$v = \begin{pmatrix} a \\ b \end{pmatrix}$$

and a is not zero, we can take

$$w = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

(what would you pick for w is a were zero?). Now, since the set $\{v, w\}$ forms a basis (this is the key idea for all n: once you know v, you need to find $n - 1$ other vectors to fill out a basis containing v) of two-dimensional space, we can find coefficients c and d so that

$$Aw = cv + dw. \tag{A.19}$$

We can summarize both (A.18) and (A.19) in one matrix equation:

$$A\,(v\,w) = (v\,w) \begin{pmatrix} \lambda & c \\ 0 & d \end{pmatrix}.$$

Here $(v\,w)$ denotes the 2 by2 matrix whose columns are the vectors v and w. To complete the construction, we let $S = (v\,w)$ and

$$\Lambda = \begin{pmatrix} \lambda & c \\ 0 & d \end{pmatrix}.$$

Then,

$$AS = S\Lambda$$

which is the same as what we wanted to prove, namely $A = S\Lambda S^{-1}$. Actually, we can even say more. It is a fundamental fact in linear algebra that, if two matrices are similar, then their eigenvalues must be the same. Now, the eigenvalues of Λ are λ and d, because the eigenvalues of any triangular matrix are its diagonal elements. Therefore, since A and Λ are similar, d must be also an eigenvalue of A.

The proof of Schur's theorem follows the same pattern, except for having fewer choices for v and w.

The Three Cases for $n = 2$

The following special cases are worth discussing in detail:

1. A has two different real eigenvalues.

2. A has two complex conjugate eigenvalues.

3. A has a repeated real eigenvalue.

In cases 1 and 2, one can always find a *diagonal* matrix Λ. To see why this is true, let us go back to the proof, but now, instead of taking just any linearly independent vector w, let us pick a special one, namely an eigenvector corresponding to the other eigenvalue of A:

$$Aw = \mu w.$$

This vector is always linearly independent of v, so the proof can be completed as before. Notice that Λ is now diagonal, because $d = \mu$ and $c = 0$.

To prove that v and w are linearly independent if they are eigenvectors for different eigenvalues, assume the contrary and show that it leads to a contradiction. Thus, suppose that $\alpha v + \beta w = 0$. Apply A to get

$$\alpha \lambda v + \beta \mu w = A(\alpha v + \beta w) = A(0) = 0.$$

On the other hand, multiplying $\alpha v + \beta w = 0$ by λ we would have $\alpha \lambda v + \beta \lambda w = 0$. Subtracting gives $\beta(\lambda - \mu)w = 0$, and as $\lambda - \mu \neq 0$ we would arrive at the conclusion that $\beta w = 0$. But w, being an eigenvector, is required to be nonzero, so we would have to have $\beta = 0$. Plugging this back into our linear dependence would give $\alpha v = 0$, which would require $\alpha = 0$ as well. This shows us that there are no nonzero coefficients α and β for which $\alpha v + \beta w = 0$, which means that the eigenvectors v and w are linearly independent.

Notice that in cases 1 and 3, the matrices Λ and S are both real. In case 1, we will interpret the solutions with initial conditions on the lines that contain v and w as "straight line solutions".

In case 2, the matrices Λ and S are, in general, not real. Note that, in case 2, if $Av = \lambda v$, taking complex conjugates gives

$$A\bar{v} = \bar{\lambda}\bar{v}$$

and we note that

$$\bar{\lambda} \neq \lambda$$

because λ is not real. So, we can always pick w to be the conjugate of v. It will turn out that solutions can be re-expressed in terms of trigonometric functions — remember example (A.13) — as we'll see in the next section.

Now let's consider Case 3 (the repeated real eigenvalue). We have that

$$\Lambda = \begin{pmatrix} \lambda & c \\ 0 & \lambda \end{pmatrix}$$

so we can also write $\Lambda = \lambda I + cN$, where N is the following matrix:

$$N = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Observe that:

$$(\lambda I + cN)^2 = (\lambda I)^2 + c^2 N^2 + 2\lambda cN = \lambda^2 I + 2\lambda cN$$

(because $N^2 = 0$) and, for the general power k, recursively:

$$\begin{aligned}
(\lambda I + cN)^k &= \left(\lambda^{k-1}I + (k-1)\lambda^{k-2}cN\right)(\lambda I + cN) \\
&= \lambda^k I + (k-1)\lambda^{k-1}cN + \lambda^{k-1}cN + (k-1)\lambda^{k-2}a^2 N^2 \\
&= \lambda^k I + k\lambda^{k-1}cN
\end{aligned}$$

so

$$(\lambda I + cN)^k t^k = \left(\lambda^k I + k\lambda^{k-1}cN\right) t^k = \begin{pmatrix} \lambda^k t^k & k\lambda^{k-1}ct^k \\ 0 & \lambda^k t^k \end{pmatrix}$$

and therefore

$$e^{\Lambda t} = \begin{pmatrix} e^{\lambda t} & cte^{\lambda t} \\ 0 & e^{\lambda t} \end{pmatrix} \tag{A.20}$$

because $0 + ct + (2\lambda c)t^2/2 + (3\lambda^2 c)t^3/6! + \cdots = cte^{\lambda t}$. (This generalizes the special case in example (A.15).)

A Shortcut

If we just want to find the form of the general solution of $Y' = AY$, we do not need to actually calculate the exponential of A and the inverse of the matrix S.

Let us first take the cases of different eigenvalues (real or complex, that is, cases 1 or 2, it doesn't matter which one). As we saw, Λ can be taken to be the diagonal matrix consisting of these eigenvalues (which we call here λ and μ instead of λ_1 and λ_2), and $S = (v \, w)$ just lists the two eigenvectors as its columns. We then know that the solution of every initial value problem $Y' = AY$, $Y(0) = Y_0$ will be of the following form:

$$Y(t) = e^{At}Y_0 = S \, e^{\Lambda t} \, S^{-1}Y_0 = (v \, w)\begin{pmatrix} e^{\lambda t} & 0 \\ 0 & e^{\mu t} \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = a \, e^{\lambda t}v + b \, e^{\mu t}w$$

where we just wrote $S^{-1}Y_0$ as a column vector of general coefficients a and b. In conclusion: *The general solution of $Y' = AY$, when A has two eigenvalues λ and μ with respective eigenvectors v and w, is of the form*

$$a \, e^{\lambda t}v + b \, e^{\mu t}w \tag{A.21}$$

for some constants a and b. So, one approach to solving IVP's is to first find eigenvalues and eigenvectors, write the solution in the above general form, and then plug-in the initial condition in order to figure out what are the right constants.

In the case of non-real eigenvalues, recall that we showed that the two eigenvalues must be conjugates of each other, and the two eigenvectors may be picked to be conjugates of each other. Let us show now that we can write (A.21) in a form which does not involve any complex numbers. In order to do so, we start by decomposing the first vector function which appears in (A.21) into its real and imaginary parts:

$$e^{\lambda t}v = Y_1(t) + iY_2(t) \tag{A.22}$$

(let us not worry for now about what the two functions Y_1 and Y_2 look like). Since μ is the conjugate of λ and w is the conjugate of v, the second term is:

$$e^{\mu t}w = Y_1(t) - iY_2(t)\,. \tag{A.23}$$

So we can write the general solution shown in (A.21) also like this:

$$a(Y_1 + iY_2) + b(Y_1 - iY_2) = (a+b)Y_1 + i(a-b)Y_2\,. \tag{A.24}$$

Now, it is easy to see that a and b must be conjugates of each other. (Do this as an optional homework problem. Use the fact that these two coefficients are the components of $S^{-1}Y_0$, and the fact that Y_0 is real and that the two columns of S are conjugates of each other.) This means that *both coefficients $a + b$ and $i(a - b)$ are real numbers.* Calling these coefficients "k_1" and "k_2", we can summarize the complex case like this: *The general solution of $Y' = AY$, when A has a non-real eigenvalue λ with respective eigenvector v, is of the form*

$$k_1\,Y_1(t)\;+\;k_2\,Y_2(t) \tag{A.25}$$

for some real constants k_1 and k_2. The functions Y_1 and Y_2 are found by the following procedure: calculate the product $e^{\lambda t}v$ and separate it into real and imaginary parts as in Equation (A.22). What do Y_1 and Y_2 really look like? This is easy to answer using Euler's formula, which gives

$$e^{\lambda t} = e^{\alpha t + i\beta t} = e^{\alpha t}(\cos \beta t + i \sin \beta t) = e^{\alpha t} \cos \beta t + ie^{\alpha t} \sin \beta t$$

where α and β are the real and imaginary parts of λ respectively.

Finally, in case 3 (repeated eigenvalues) we can write, instead:

$$Y(t) = e^{At}Y_0 = S\,e^{\Lambda t}\,S^{-1}Y_0 \;=\; (v\;w)\begin{pmatrix} e^{\lambda t} & cte^{\lambda t} \\ 0 & e^{\lambda t} \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix}$$
$$=\; a\,e^{\lambda t}v + b\,e^{\lambda t}(ctv + w)\,.$$

When $c = 0$ we have from $A = S\Lambda S^{-1}$ that A must have been the diagonal matrix

$$\begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}$$

to start with (because S and Λ commute). When $c \neq 0$, we can write $k_2 = bc$ and redefine w as $\frac{1}{c}w$. Note that then (A.19) becomes $Aw = v + \lambda w$, that is, $(A - \lambda I)w = v$. Any vector w with this property is linearly independent from v (why?).

So we conclude, for the case of repeated eigenvalues: *The general solution of $Y' = AY$, when A has a repeated (real) eigenvalue λ is either of the form $e^{\lambda t}Y_0$ (if A is a diagonal matrix) or, otherwise, is of the form*

$$k_1\,e^{\lambda t}v\;+\;k_2\,e^{\lambda t}(tv + w) \tag{A.26}$$

for some real constants k_1 and k_2, where v is an eigenvector corresponding to λ and w is any vector which satisfies $(A - \lambda I)w = v$. Observe that $(A - \lambda I)^2 w = (A - \lambda I)v = 0$. general, one calls any nonzero vector such that $(A - \lambda I)^k w = 0$ a *generalized eigenvector* (of order k) of the matrix A (since, when $k = 1$, we have eigenvectors).

Forcing Terms

The use of matrix exponentials also helps explain much of what is done in chapter 4 (forced systems), and renders Laplace transforms unnecessary. Let us consider non-homogeneous linear differential equations of this type:

$$\frac{dY}{dt}(t) = AY(t) + u(t). \tag{A.27}$$

We wrote the arguments "t" just this one time, to emphasize that everything is a function of t, but from now on we will drop the t's when they are clear from the context.

Let us write, *just as we did when discussing scalar linear equations*, $Y' - AY = u$. We consider the "integrating factor" $M(t) = e^{-At}$. Multiplying both sides of the equation by M, we have, since $(e^{-At}Y)' = e^{-At}Y' - e^{-At}AY$ (*right?*):

$$\frac{de^{-At}Y}{dt} = e^{-At}u.$$

Taking antiderivatives:

$$e^{-At}Y = \int_0^t e^{-As}u(s)\,ds + Y_0$$

for some constant vector Y_0. Finally, multiplying by e^{-At} and remembering that $e^{-At}e^{At} = I$, we conclude:

$$Y(t) = e^{At}Y_0 + e^{At}\int_0^t e^{-As}u(s)\,ds. \tag{A.28}$$

This is sometimes called the "variation of parameters" form of the general solution of the forced equation (A.27). Of course, $Y_0 = Y(0)$ (just plug-in $t = 0$ on both sides).

Notice that, if the vector function $u(t)$ is a polynomial in t, then the integral in (A.28) will be a combination of exponentials and powers of t (integrate by parts). Similarly, if $u(t)$ is a combination of trigonometric functions, the integral will also combine trigonometric functions and polynomials. This observation justifies the "guesses" made for forced systems in chapter 4 (they are, of course, not guesses, but consequences of integration by parts).

Exercises

1. In each of the following, factor the matrix A into a product $S\Lambda S^{-1}$, with Λ diagonal:

a. $\quad A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$

b. $\quad A = \begin{pmatrix} 5 & 6 \\ -1 & -2 \end{pmatrix}$

c. $\quad A = \begin{pmatrix} 2 & -8 \\ 1 & -4 \end{pmatrix}$

d. $\quad A = \begin{pmatrix} 2 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & -1 \end{pmatrix}$

2. For each of the matrices in Exercise 1, use the $S\Lambda S^{-1}$ factorization to calculate A^6 (do *not* just multiply A by itself).

3. For each of the matrices in Exercise 1, use the $S\Lambda S^{-1}$ factorization to calculate e^{At}.

4. Calculate e^{At} for this matrix:

$$\begin{pmatrix} 0 & 1 & 2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

   using the power series definition.

5. Consider these matrices:

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \qquad B = \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix}$$

   and calculate e^{At}, e^{Bt}, and $e^{(A+B)t}$.

   Answer, true or false: is $e^{At}e^{Bt} = e^{(A+B)t}$?

6. (Challenge problem) Show that, for any two matrices A and B, it is true that

$$e^{At}e^{Bt} = e^{(A+B)t} \quad \text{for all } t$$

   if and only if $AB - BA = 0$. (The expression "$AB - BA$" is called the "Lie bracket" of the two matrices A and B, and it plays a central role in the advanced theory of differential equations.)

Index