School of Mathematical Sciences

Mathematical Biology (Honours)

Cell and tissue, shell and bone, leaf and flower, are so many portions of matter, and it is in obedience to the laws of physics that their particles have been moved, moulded and conformed. They are no exception to the rule that God always geometrizes. Their problems of form are in the first instance mathematical problems, their problems of growth are essentially physical problems, and the morphologist is, ipso facto, a student of physical science.

-Sir D'Arcy Wentworth Thompson, *'On Growth and Form'* [6]

---

1 Introduction: Basic Ideas in Mathematical Modelling

1.1 Mathematical modelling

The aim of mathematical modelling is to improve our understanding of some real-life problem. In areas such as physics, chemistry or engineering, mathematics has been so important in testing hypotheses and making predictions that (to the dismay of many students!) it is now impossible to study these subjects without first learning a significant amount of mathematics. This is not true to the same extent in biology, but their conspicuous success in the physical sciences has meant that there is an increasing readiness to apply mathematical techniques to biological problems. Correspondingly, the field of mathematical biology has grown rapidly in the last 20 years or so.

The general process of mathematical modelling can be summarised as follows:

1. Develop a hypothesis regarding the important mechanisms underlying the problem. This requires a detailed knowledge of the processes at work. In mathematical biology research, the input of experimental collaborators is often very important at this stage.

2. Formulate the hypothesis in mathematical terms. The often involves recasting the assumptions you have made about what is happening in the problem in the form of differential equations.

3. Analyse the resulting mathematical problem so as to understand the behaviour of the solution. (For very simple models, you may be able to write down a closed-form solution; in other cases, more complicated techniques will need to be applied to gain insights into its behaviour.)

4. Interpret the mathematical results in the context of the original problem. Is the solution consistent with what is observed in experiments? If so, are there new experiments you could devise based on your model, which would further test its validity? If not, what assumptions in your model might you need to revisit?

As mathematicians, your training up to now has probably focused mainly on stage 3; however, in many cases stage 2 can be equally if not more important. A problem which has been experimentally intractable to biologists can sometimes be solved quite straightforwardly once it has been formulated in mathematical terms. Similarly, a mathematical formula on a piece of paper is not likely to be much use to a biologist; you need to see the significance (and the limitations) of the result in the context of the problem (stage 4) if any real biological insight is to be gained. One of the aims of this course is to help you develop these additional skills.

1.2 Building mathematical models

One of the golden rules of mathematical modelling is: **keep it simple!** No understanding will be gained by converting an intractable biological problem into a mathematical model that is too complicated to analyse. When building a model, we must to focus on the processes we think are the most important, and neglect the others, at least to begin with. (Once the basic model is fully understood, additional effects can be added to it and their effect on the solution investigated.) Knowing what to put in, and what to leave out, is something of an art, and requires experience. However, there are a couple of basic techniques that can be very helpful.

1.2.1 Dimensional analysis

It is obvious that in order to be physically consistent, we can only equate together quantities which have the same dimensions - *e.g.* in Newton's Second Law, $F = ma$, the quantities on both sides of the equation must have the dimensions of force. Measurements of quantities are taken with respect to a reference value; the particular reference value used will depend upon the system of units adopted. Physical relationships must be true irrespective of the system of units used. In defining a system of units some quantities are considered fundamental - *e.g.* mass, length and time, which we shall denote [M], [L] and [T]. Other units are derived from these - *e.g.* speed is the rate of change of distance, and so has dimensions of [L] $[\mathrm{T}]^{-1}$. Other fundamental units include electric charge, [Q], and temperature, [Θ].

Example: Coulomb's law
The force, F acting on two particles with charges q_1 and q_2, separated by a distance r, is given by

$$F = k_e \frac{q_1 q_2}{r^2}$$

where k_e is a constant. What are the dimensions of k_e?

Both sides of the equation must have the dimensions of force - *i.e.* [M] [L] $[T]^{-2}$. Hence

$$\frac{[M][L]}{[T]^2} = [k_e]\frac{[Q]^2}{[L]^2}, \qquad \Rightarrow \qquad [k_e] = \frac{[M][L]^3}{[Q]^2[T]^2}.$$

A quantity which has no units is said to be **dimensionless**. Recall that the length, a, of an arc of a circle of radius r, subtended by an angle, θ, is given by $a = r\theta$. Since both a and r have dimensions of [L], $[\theta] = [1]$ (*i.e.* θ is dimensionless). Note that we can create dimensionless quantities by making appropriate combinations of dimensional quantities. For example, in the case of a simple pendulum of length, l, which oscillates with a frequency ω in a gravitational field of strength, g, the quantity $\dfrac{g}{l\omega^2}$ is dimensionless (exercise).

We can exploit dimensionless quantities to help us deduce model equations. Suppose we are considering a problem with dimensional variables x_1, $x_2, \ldots, x_n$ and dimensional parameters (constants) $\alpha_1, \ldots, \alpha_m$, and want to determine the relationship between these quantities. Since we can only equate quantities of the same dimension, this limits the combinations of the x_i and α_j which are possible. Obviously, dimensionless quantities can be equated in this way, so if the x_i and α_j can be combined to create k dimensionless groups, β_1, $\beta_2, \ldots, \beta_k$, the relationship between these can be written in the format

$$f(\beta_1, \beta_2, \ldots, \beta_k) = 0.$$

It is frequently the case that only a small number of dimensionless groups can be created from the dimensional variables and parameters which we believe to influence the problem. This will simplify the modelling very considerably. For example, if there is only one dimensionless group, we have must have a relationship of the form $f(\beta) = 0$. This means β is a zero of the function f - *i.e.* the physical relationship between our variables can be expressed as $\beta = constant$.

Example: How powerful is an atomic bomb?

In 1945, the USA test-detonated the world's first atomic bomb, a device code-named Trinity, in the New Mexico desert. Information concerning the test was highly classified by the US government at the time (in fact, the full technical report on the explosion was not published until 1976), though a series of photographs of the explosion (which included the time since detonation) were declassified in 1947. The UK government was intensely interested in the results of the US test, as they wanted to develop their own atomic weapons; in particular, they wanted to know how much energy such a device might release. They asked Sir Geoffrey Taylor (G. I. Taylor) an applied mathematician at the University of Cambridge, to work on the problem. For explosions which took place in the open, he assumed the blast wave created would be spherical in shape. He reasoned that, in the early stages of an explosion (before energy could be radiated as heat), the radius of the blast, R, could only depend on the energy of the explosion, E, the time since detonation, t, and the density of the air, ρ. By using dimensional analysis, he realised that only one dimensionless number could be created from these quantities, and so

$$\frac{Et^2}{\rho R^5} = constant.$$

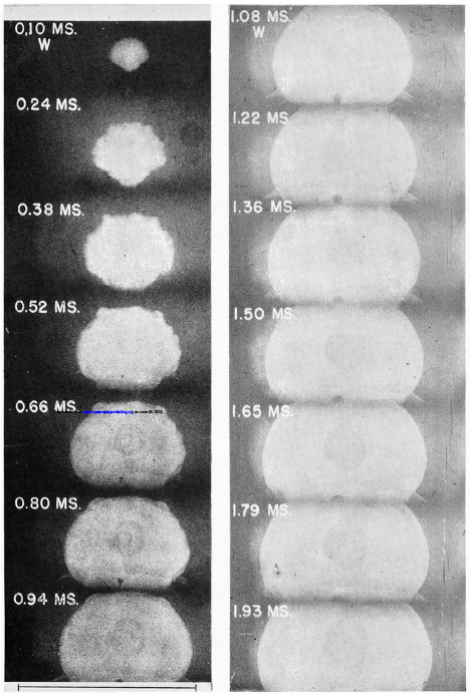

Figure 1: Photographs of the Trinity atomic bomb test used by G. I. Taylor (from [5]).

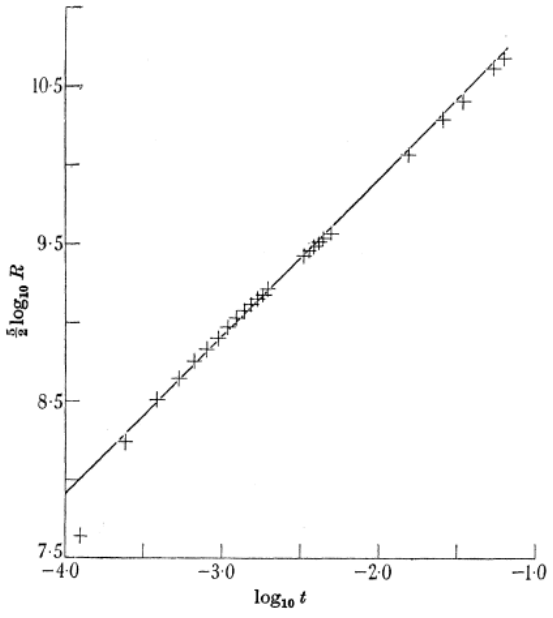

Figure 2: Graph of data from the photographs, showing clearly that $R^5 \propto t^2$ (from [5]).

By means of other arguments, Taylor determined that the value of the constant would be approximately one.

Assuming a typical value for the density of air, ρ, and making measurements from the published photographs of the test to estimate R at each value of t, Taylor was able to obtain a value for E, the energy of the blast, of $16.8 - 23.7$ kT (*i.e.* equivalent to 16,800-23,700 tons of TNT). The official test determined it was 20 kT. Taylor published his results in 1950 [5], two years before the first UK atomic weapons test.

1.2.2 Nondimensionalisation and scaling

Although dimensional analysis can sometimes help us to deduce the equations governing a process, another important reason for dealing with dimensionless quantities is that it allows us to compare the relative importance of various effects very easily. This is not really true with dimensional equations.

Example: Suppose for a moment, we are interested in the spread of a pollutant being released from a factory chimney into the air over a small town. The chemical will diffuse, and will also be carried on the wind. We are told the typical wind speed in the area is 5 km h^{-1}, and the diffusion coefficient of the chemical in air at 25 °C has been measured as 0.3 cm^2 s^{-1}. Is the pollutant transported mainly by diffusion, or on the wind?

In order to answer this question, we first need to think about how the pollutant concentration, c evolves. For simplicity, let us just consider the spread of the pollutant in one dimension, and assume that the concentration profile is steady (*i.e.* does not change with time). We let x be the distance downwind of the chimney. Then the concentration obeys

$$U\frac{\partial c}{\partial x} = D\frac{\partial^2 c}{\partial x^2}, \qquad c = c_0 \quad \text{at } x = 0, \qquad c \to 0 \quad \text{as } x \to \infty,$$

where U is the wind speed and D is the diffusion coefficient. (You can just accept this equation for now, we will derive it later in the course.) We are interested in how the pollutant spreads over a town, so a typical lengthscale (distance) we be around a kilometre; let L be the distance from the chimney to the town centre. Then we notice there are some obvious natural scales in the problem. It is convenient to specify the concentration in terms of the fraction of the value at the site of release, c_0, and similarly to measure length in terms of the fraction of the distance to the town centre. Hence we set

$$\tilde{x} = \frac{x}{L}, \qquad \tilde{c} = \frac{c}{c_0},$$

where the tildes indicate dimensionless quantities. We then note that UL/D is a dimensionless quantity, which is called the Péclet number, $\mathcal{P}$. Hence the dimensionless concentration obeys

$$\mathcal{P}\frac{\partial \tilde{c}}{\partial \tilde{x}} = \frac{\partial^2 \tilde{c}}{\partial \tilde{x}^2}, \qquad \tilde{c} = 1 \quad \text{at } x = 0, \qquad \tilde{c} \to 0 \quad \text{as } \tilde{x} \to \infty.$$

This equation has some obvious superficial advantages over the dimensional version: there are fewer constants, so there is less chance of us making a mistake. But more

importantly, we now see that if $\mathcal{P} \ll 1$, diffusion is the dominant transport mechanism, whilst for $\mathcal{P} \gg 1$, transport on the wind is the more important factor. If $\mathcal{P} = O(1)$, then both mechanisms play an equal role. For our situation, we find that $U \approx 1.4$ m s^{-1}, $D = 3 \times 10^{-5}$ m^2 s^{-1} and $L = 10^3$ m. Hence

$$\mathcal{P} \approx \frac{1.4 \times 1000}{3 \times 10^{-5}} \approx 5 \times 10^7 \gg 1.$$

Therefore, transport by the wind is vastly more important than diffusion, so we can make our lives easier by neglecting the diffusion term in the equation.

To summarise, it is very helpful to work with dimensionless models for the following reasons:

- The equations involve fewer symbols, so we are less likely to make mistakes in calculations, and it is often easier to recognise the type of equations involved. Since the coefficients are real numbers, rather than dimensional quantities, their magnitudes can be directly compared, which is useful for determining the most important effects in the problem, and making simplifications where appropriate.
- Reducing the number of parameters means results can be investigated more quickly and presented in more compact form. Above, we reduced the number of parameters from 3 (c_0, U and D) to one. Hence the behaviour of the solution depends on only one parameter, $\mathcal{P}$ - we do not have to give separate plots for different values of c_0, U and D). This can be particularly important when we have to solve a problem numerically - especially if the simulation takes a long time to run.
- Solutions obtained for one system can be applied to another which obeys the same equation, but with different parameter values - there is no need to recalculate the solution.
- Often the dimensionless equations can help in the design of experiments - *e.g.* we could investigate the pollutant dispersal problem experimentally by building a scale model in the lab, such that $\mathcal{P}$ takes the same value as for the real problem.

1.3 Types of models

In this course, we will concentrate on models based on ordinary and partial differential equations. As these are so numerous in the literature, it might be tempting to think that they are a good way to attack any real-life problem: this is certainly not the case (we will discuss the cautionary example from [1] in class). It is useful to remind ourselves at this point that there may be other more appropriate modelling approaches for certain problems, such as:

- **Difference equations** (also known as recurrence relations, iterated maps, or just maps): The dependent variable can be discrete or continuous; time is always discrete; suitable for seasonal events; can have deterministic and stochastic difference equations.

- **Stochastic processes:** A family of random variables $\{X(t)\}$, indexed by a parameter t, is called a stochastic process; Markov-chain models are one class of (memory-less) stochastic model; particularly useful for small populations.
- **Cellular automata:** Fully discrete models, all independent variables and all dependent variables are discrete; analysis is mainly restricted to computer analysis and numerical simulation; can be either deterministic or stochastic, using a random number generator.

2 Biochemical reactions

In the next two sections, we will consider mathematical models consisting of ordinary differential equations, with which you will already be familiar. As much of the challenge in mathematical modelling is simply knowing what equations to write down, we give considerable attention to this aspect. We consider how the important variables are identified, and how we can use the processes of scaling and inspectional analysis introduced earlier to identify appropriate simplifying assumptions. Since only the simplest models can be solved analytically, we also introduce techniques which allow us to gain insight into the qualitative behaviour of the models, without actually having to solve them.

2.1 Chemical reactions

Chemical reactions are important in many biological processes - *e.g.* digestion, photosynthesis, respiration. We first consider an irreversible reaction process in which reactants A and B produce C - *i.e.*

$$A + B \xrightarrow{k} C, \quad k = \text{the reaction constant.}$$

Let $a = [A]$, $b = [B]$, $c = [C]$ denote the concentrations of A, B and C, respectively. (The use of square brackets to indicate a concentration is standard in biology and chemistry, and hence appears in many papers and books. However, I think it can result in equations that look very cluttered, and so will try to avoid it in this course.) The SI units of concentration are moles m^{-3} (abbreviated to mol m^{-3}), where a *mole* is simply a number of molecules - 6.023×10^{23}. Since a cubic metre is often an inconveniently large volume, moles per litre (also called *molar*, M) is a common alternative unit.

We argue

$$\left\{ \begin{array}{c} \text{change of} \\ \text{the product} \\ \text{over time} \end{array} \right\} = \left\{ \begin{array}{c} \text{number of} \\ \text{collisions of} \\ \text{molecules } A \\ \text{and } B \end{array} \right\} \cdot \left\{ \begin{array}{c} \text{probability that a} \\ \text{collision has enough} \\ \text{kinetic energy to} \\ \text{initiate a reaction} \end{array} \right\},$$

which yields the ODE

$$\frac{dc}{dt} = k\, a\, b.$$

This is the **Law of Mass Action**. While called a law, it is really just a mathematical model, which is useful in many, but not all situations. Reactions which obey mass action kinetics are called *elementary reactions*. However, there are many biological reactions which proceed through complex mechanisms consisting of several elementary steps which are not known in sufficient detail for the law of mass action to be applied.

It is also worth remarking that the assumption that the reaction rate, k, is constant is strictly only true if the reaction occurs at a constant temperature. Many chemical

reactions give off, or take in, significant amounts of heat. (In modelling such a case, we would need to know the reaction rate as a function of temperature, and either measure, or have a model for, the temperature throughout the reaction. This would make the modelling much more complicated.) However, for reactions occurring under physiological conditions, the assumption of constant temperature is usually quite appropriate - *e.g.* body temperature in mammals is maintained very close to constant under a wide range of external conditions.

Extension: What equation would we write down for a reaction involving m molecules of A and n molecules of B reacting to produce a molecule of C?

Next, consider a reversible reaction

$$A + B \underset{k_-}{\overset{k_+}{\rightleftharpoons}} C.$$

Here, as for many biological processes it is necessary to follow the time evolution of more than one factor, leading to systems of ODEs. Balancing the production and consumption terms for each participating chemical species gives

$$\begin{aligned} \frac{dc}{dt} &= k_+ a\, b - k_- c, \\ \frac{da}{dt} &= -k_+ a\, b + k_- c, \\ \frac{db}{dt} &= -k_+ a\, b + k_- c. \end{aligned}$$

Note that this system is reducible to just one equation, since

$$\frac{d}{dt}(c + a) = \frac{d}{dt}(c + b) = 0,$$

and given initial concentrations a_i, b_i, c_i we have $a = c_i + a_i - c$ and $b = c_i + b_i - c$.

2.2 Enzymes

A *catalyst* is a substance that speeds up a chemical reaction without itself being consumed in the reaction. An *enzyme* is a biological catalyst: most are proteins. The increase in the rate of reaction they induce is staggering - often of the order of several million-fold. The substances on which they operate are called *substrates*.

Enzymes regulate a vast array of processes in the body, particularly those related to digestion and metabolism. For example, amylase in saliva starts to break down starch in our food into sugars, which are in turn further broken down into glucose, the body's primary source of energy. The release of energy from glucose also occurs through a chain of enzyme-mediated reactions. However, they are also important industrially. An everyday example would be the use of enzymes in laundry powder, which help to

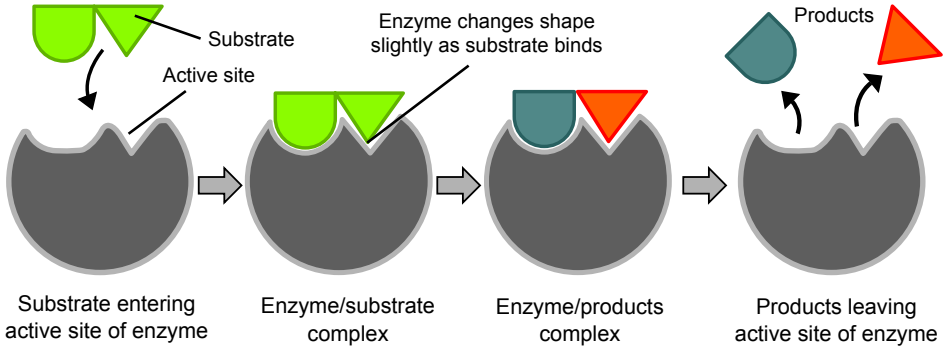

Figure 3: Schematic of the induced fit theory of enzyme action. (Diagram credit: TimVickers, Wikimedia.)

break down the fats in stains.

Each enzyme catalyses a specific reaction. This specificity is usually explained in terms of the 'lock and key' theory (or its slight modification- the induced fit theory, shown diagrammatically in Fig. 3). The idea is that the enzyme and its substrates both have specific, complementary shapes. The substrate must bind to the enzyme in order for the reaction to occur, and this is only possible when the two fit together like a 'lock and key'. (The induced fit theory allows for some flexibility in the enzyme, which helps the binding become tighter.) Thus, the enzyme is unable to bond with other substrates, as they do not have the correct shape.

Since enzymes are proteins, they tend to work best within a narrow range of temperature and pH. At low temperatures, molecules have low kinetic energy, so the number of collisions is reduced, and hence the rate of reaction. At temperatures above around $40°C$ (or extremes of pH), proteins become *denatured*, meaning their structure is changed (*e.g.* think of what happens to egg white as it is cooked). This makes the enzyme less effective at catalysing the reaction.

2.3 Michaelis-Menten kinetics

Consider the enzymatic reaction

$$E + S \underset{k_-}{\overset{k_+}{\rightleftharpoons}} E + P,$$

involving the reaction of a substrate S with an enzyme E to gain a product P, and the enzyme. The enzyme is a catalyst.

To allow use of the Law of Mass Action, we assume E and S form an intermediate complex C which then decays into P and E:

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} C \overset{k_2}{\rightarrow} E + P, \quad \text{the } \textit{Michaelis-Menten scheme.}$$

Letting $s = [S]$, $e = [E]$, $c = [C]$, $p = [P]$, we can describe the process using ODEs:

$$\begin{aligned}
\frac{ds}{dt} &= -k_1 s\, e + k_{-1} c, \\
\frac{de}{dt} &= -k_1 s\, e + k_{-1} c + k_2 c, \\
\frac{dc}{dt} &= k_1 s\, e - k_{-1} c - k_2 c, \\
\frac{dp}{dt} &= k_2 c.
\end{aligned}$$

Note that the units of the constant k_1 differ from those of k_-1 and k_2. We take the initial conditions for the system to be

$$s(0) = s_i > 0, \qquad e(0) = e_i > 0, \qquad c(0) = p(0) = 0. \tag{2.1}$$

We note that by adding the second and third equations above, integrating and applying the initial conditions, we find that

$$e(t) + c(t) = e_i. \tag{2.2}$$

Hence we can eliminate e from the equations for s and c. Since p depends only on c, we need only consider the reduced system

$$\frac{ds}{dt} = -k_1 s\,(e_i - c) + k_{-1} c = -k_1 s e_i + (k_1 s + k_{-1}) c, \tag{2.3a}$$

$$\frac{dc}{dt} = k_1 s\,(e_i - c) - k_{-1} c - k_2 c = k_1 s e_i - (k_1 s + k_{-1} + k_2) c, \tag{2.3b}$$

with initial conditions $s(0) = s_i$, $c(0) = 0$.

These have no known exact solution but approximate solutions have been obtained. The best known and most commonly used is the approximation introduced by Michaelis and Menten in 1913, which is motivated by the observation that the availability of the enzyme is often the limiting factor in the reaction. In order to develop our approximate solution, we must use the techniques of scaling and inspectional analysis we met earlier in the course.

We translate the assumption that the reaction is limited by the availability of the enzyme into the mathematical assumption that $e_i \ll s_i$. We let T be the timescale over which the substrate is observed to be converted into product (we will specify this in terms of the system parameters shortly). Since we start with a finite amount of substrate and no product, we must have $0 \le s \le s_i$ and $0 \le p \le s_i$. From (2.2) we have $0 \le c \le e_i$, so $c \ll s_i$. We hence nondimensionalise our variables as follows (where tildes indicate dimensionless variables):

$$t = T\tilde{t}, \qquad s = s_i \tilde{s}, \qquad c = \epsilon s_i \tilde{c},$$

where we have introduced the dimensionless parameter $\epsilon = e_i/s_i \ll 1$. Equation (2.3a) then becomes (dropping tildes):

$$\frac{ds}{dt} = \epsilon k_1 s_i T(cs - s) + \epsilon k_{-1} T c, \tag{2.4}$$

Since, by definition, T is the timescale for s to undergo an $O(1)$ change, we expect the parameters on the RHS of (2.4) to be $O(1)$. Hence we choose $T = 1/\epsilon k_1 s_i$, which implies that the timescale on which the substrate is converted to product is much longer than the timescale of complex formation. We also assume $A = \epsilon k_{-1} T = k_{-1}/k_1 s_i = O(1)$, so the rates of complex formation and dissociation are of the same order of magnitude. Hence (2.3) simplifies to

$$\frac{ds}{dt} = cs - s + Ac, \tag{2.5a}$$

$$\epsilon \frac{dc}{dt} = s - cs - (A + B)c, \tag{2.5b}$$

where $B = k_2/k_1 s_i$ is the ratio of the rate of product formation to complex formation, and the initial conditions become $s(0) = 1$, $c(0) = 0$.

We now expand s and c as power series in the small parameter, ϵ, so

$$s = s_0 + \epsilon s_1 + O(\epsilon^2), \qquad c = c_0 + \epsilon c_1 + O(\epsilon^2).$$

Substituting the above into (2.5), at leading order we find

$$c_0 = \frac{s_0}{K_m + s_0} \qquad \text{where} \quad K_m = A + B = \frac{k_{-1} + k_2}{k_1 s_i} \tag{2.6}$$

$$\frac{ds_0}{dt} = \frac{-B s_0}{K_m + s_0}, . \tag{2.7}$$

This is the most common Michaelis-Menten function used, among other things, for rate of depletion of nutrients in biological systems. B is the maximum uptake rate; K_m is the concentration at which the uptake rate is one half of the maximum.

Remarks:

- Note how the different timescales in the reaction have emerged naturally from the scaling analysis. In fact, it is possible to show that the approximate solution remains valid under the weaker assumption that $\epsilon/(K_m + 1) \ll 1$ (*i.e.* either there is little enzyme compared to substrate, or the rates of complex dissociation and product formation are slow compared to the rate of complex formation). Full details are given in [4].

- Our leading order solution for c, c_0 as given in equation (2.6) does not obey the initial condition $c(0) = 0$. This is because, in obtaining it, we neglected the time derivative in equation (2.5b). Situations like this, where the small parameter multiplies the highest derivative in an equation are called **singular perturbation problems**, and generally give rise to a leading-order solution that does not obey one or more of the initial or boundary conditions. If, in equation (2.5b) we introduce the new, shorter timescale $t = \epsilon\tau$, then the time derivative term becomes $O(1)$. Physically, this implies that there is a short timescale in the problem on which c changes from zero to the value given in (2.6). See Chapter 6 of [3] for full details.

3 Excitable systems

3.1 Background

The cell membrane is a *phospholipid bilayer* separating the cell interior (the cytoplasm) from the extracellular environment. The membrane contains numerous proteins, and is approximately 7.5nm thick. The most important property of the cell membrane is its selective permeability: it allows the passage of some molecules but restricts the passage of others, thereby regulating the passage of materials into and out of the cell. Many substances penetrate the cell membrane at rates reflected by their diffusive behaviour in a pure phospholipid bilayer. However, certain molecules and ions such as glucose, amino acids and Na^+ pass through cell membranes much more rapidly, indicating that the membrane proteins selectively facilitate transport.

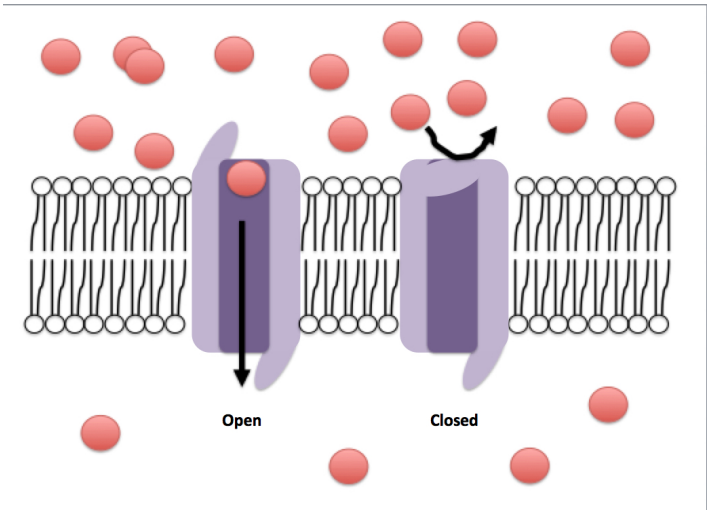

Figure 4: Illustration of a channel in the cell membrane (Credit: Efazzari [CC BY-SA 4.0])

The membrane contains water-filled pores with diameters of about 0.8nm, and protein lined pores, called channels or gates, which allow the passage of specific molecules. Both the intracellular and extracellular environments comprise (among other things) a dilute aqueous solution of dissolved salts, mainly NaCl and KCl, which dissociate into Na^+, K^+ and Cl^- ions. The cell membrane acts as a barrier to the free flow of these ions and to the flow of water.

The mechanisms that facilitate transport across the cellular membrane can be divided into active and passive processes. Active processes requires energy expenditure, while passive processes result solely from the random motion of molecules, for example, diffusion.

Action potentials, or 'nerve impulses', are brief changes in the membrane potential of a cell produced by the flow of ionic current across the cell membrane. They enable communication by many cell types, including neurons, cardiac and muscle cells.

First we note that:

- Numerous fundamental particles, ions and molecules have an electric charge, *e.g.* the electron, e^-, and the sodium ion, Na^+. The SI unit of electrical charge is the Coulomb (C).
- It is an empirical fact that total charge is conserved.
- Electric charges exert electrical forces on one another such that like charges repel and unlike charges attract. The electric potential, denoted V, is the potential energy of a unit of charge due to such forces and is measured in volts (V) or Joules per Coulomb (JC^{-1}).
- A concentration of positive particles has a large *positive* potential, while a concentration of negative particles has a large, but *negative* potential.
- Electric current is defined to be the rate of flow of electric charge, measured in Amperes, A (also known as Amps; equivalently, Cs^{-1}).

3.1.1 Revision of electrical circuits

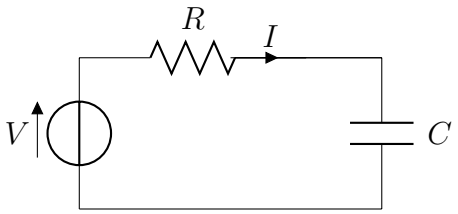

Figure 5: A simple electrical circuit with a voltage source, resistor and capacitor

Before we consider modelling electrical behaviour in cells, recall the simple electrical circuit shown in Figure 5 where:

- $q(t)$ is the charge;
- $I(t) = \dfrac{dq}{dt}$ is the current (rate of flow of charge);
- $V(t)$ is the potential difference / voltage (which causes the movement of charge - somewhat analogous to pressure in fluid mechanics);
- R is the resistance (measured in Ohms, Ω);
- $g = \dfrac{1}{R}$ is the conductance;
- C is the capacitance (ability to store charge) - measured in Farads, F.

We have:

- **Ohm's law*** - the potential difference (voltage drop) across a resistor is proportional to the current through the resistor.

$$V_R(t) = I(t)R = \frac{I(t)}{g}.$$

- **Faraday's law** - the potential difference across a capacitor is proportional to the charge stored.

$$V_c(t) = \frac{q(t)}{C}.$$

- **Kirchoff's law** - the voltage supplied is equal to the total voltage drop

$$V(t) = V_R(t) + V_C(t).$$

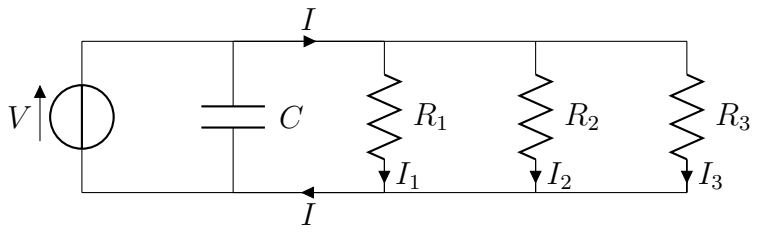

Figure 6: A more complex electrical circuit with a voltage source, three resistors and a capacitor

- For elements in parallel (see Figure 6), total current is the sum of the currents in each branch, whilst the potential difference across each is the same - *e.g.*

$$I(t) = I_1(t) + I_2(t) + I_3(t) = g_1V + g_2V + g_3V.$$

Now since $V(t) = q/C$

$$\frac{dV}{dt} = \frac{1}{C}\frac{dq}{dt} = \frac{I}{C},$$

and so

$$\frac{dV}{dt} = \frac{V}{C}(g_1 + g_2 + g_3).$$

In fact, when we are thinking about ionic currents across the cell membrane, things are a little more complicated than in electrical circuits. Instead of the current being directly proportional to the potential ($I = gV$), we usually have a relationship of the form

$$I = g(V)(V - V^*),$$

where $g(V)$ is the ion-specific membrane conductance, and V^* is the *Nernst potential*. We will explain in more detail how they arise in the next two sections.

*This 'law', as for many others, is in fact just a model.

3.2 The membrane potential

Active and passive transport of ions into and out of the cell can result in differences in ion concentrations between the interior and exterior of a cell. This induces a potential difference across the membrane, and in turn affects ion transport.

Suppose we have two reservoirs containing different concentrations of a positively charged ion X^+. We suppose that both reservoirs are electrically neutral to begin with, so that there is an equal concentration of a negatively charged ion Y^-. Now suppose that the reservoirs are separated by a semi-permeable membrane which is permeable to X^+ but not to Y^-. Then the difference in concentration of X^+ on each side will lead to the flow of X^+ across the membrane. However, because Y^- cannot diffuse through the membrane this will lead to a build up of charge on one side. This charge imbalance sets up an electric field, which produces a force on the ions opposing further diffusion of X^+. (The actual amount of X^+ which diffuses through the membrane is small, and the excess charge all accumulates near the interface, so that to a goodap-proximation the solutions on either side remain electrically neutral.) The potential difference at which equilibrium is established and diffusion and electric-field-generated fluxes balance is known as the *Nernst potential*.

Instead of using the Ohm's law relation $I = gV$ for the ionic current across the cell membrane, it is common to assume a relationship of the form

$$I = g(V - V^*),$$

V^* is the *Nernst potential*. This gives an Ohm's law-like linear dependence of current on V, but takes into account the fact that there is no net flux of ions across the membrane when $V = V^*$.

3.3 Gating

It is found experimentally that the conductance g is not constant but depends on both V and time t. One proposed explanation for this is that the channels are not always open, but may be open or closed, and that the transition rates between open and closed states depends on the potential difference, V . The membrane conductance may then be written as ng, where g is the constant conductance that would result if all the channels were open, and n is the proportion of open channels.

3.3.1 Simple gates

Consider a generic ion, with n being the proportion of open ion channels. Denoting the open channels by O and the closed channels by C, the reaction scheme is simply

$$C \underset{\beta(V)}{\overset{\alpha(V)}{\rightleftharpoons}} O.$$

where $\alpha(V)$ and $\beta(V)$ represent voltage dependent rates of switching between the closed and open states. Using the law of mass action we obtain

$$\frac{dn}{dt} = \alpha(V)(1-n) - \beta(V)n,$$

or equivalently,

$$\tau_n(V)\frac{dn}{dt} = n_\infty(V) - n,$$

where $n_\infty(V) = \alpha/(\alpha+\beta)$is the equilibrium value of n and $\tau_n(V) = 1/(\alpha+\beta)$ is the timescale for approach to this equilibrium (both of which can be determined experimentally).

3.3.2 Multiple gates

We can now extend the simple model above to channels composed of multiple subunits (gates), each one of which can be in either the open or closed state.

We start by assuming that the channel consists of two gates, which may both exist in open or closed states. The ion channel is open only if both gates are open; the ion channel is closed if any one gate within the ion channel is closed. Let S_i $(i = 0, 1, 2)$ denote the proportion of channels with i gates open. Then, the reaction scheme is

$$S_0 \underset{\beta(V)}{\overset{2\alpha(V)}{\rightleftharpoons}} S_1 \underset{2\beta(V)}{\overset{\alpha(V)}{\rightleftharpoons}} S_2,$$

where the factors of two arise because there are two possible states with one gate open and one gate closed (since each gate is identical we lump these two states into one variable S_1). Note we also have the overall constraint that

$$S_0 + S_1 + s_2 = 1. \tag{3.1a}$$

Using mass action kinetics we have

$$\frac{dS_0}{dt} = \beta(V)(1 - S_0 - S_2) - 2\alpha(V)S_0, \tag{3.1b}$$

$$\frac{dS_2}{dt} = \alpha(V)(1 - S_0 - S_2) - 2\beta(V)S_2, \tag{3.1c}$$

where, in general, V could itself be a function of time (and space). Note, we could have written down a third equation for S_1 but this is not needed, since it can be found using (3.1a).

Now, the proportion of open gates n is given by

$$n = \frac{1}{2}S_1 + S_2 = \frac{1}{2}(1 - S_0 + S_2),$$

so, by taking the appropriate linear combination of the equation for S_0 and S_1 we find that

$$\frac{dn}{dt} = \alpha(V)(1-n) - \beta(V)n. \tag{3.2}$$

The system (3.1) is satisfied by

$$S_0 = (1-n)^2, \quad S_1 = 2n(1-n), \quad S_2 = n^2, \tag{3.3}$$

(which can be verified by simple substitution). In fact, by writing $S_0 = (1-n)^2 + z_0$, $S_2 = n^2 + z_2$ (so that $S_1 = 2n(1-n) - z_0 - z_2$) and substituting in equation (3.1b) and (3.1c) we find

$$\frac{dz_0}{dt} = -2\alpha z_0 - \beta(z_0 + z_2), \qquad \frac{dz_2}{dt} = -\alpha(z_0 + z_2) - 2\beta z_2.$$

This is a linear system with eigenvalues $-(\alpha+\beta)$ and $-2(\alpha + beta)$, so z_0 and z_2 will decay exponentially to zero. Hence, the solution of the system will always approach exponentially that given by equations (3.2) and (3.3).

The analysis of a two-gated channel generalised easily to channels containing more gates. In the case of k identical gates the fraction of open channels is n^k, where n again satisfies (3.2). It has been found that a model with 4 gates agrees with empirical observations of K^+ channels, a fact we will exploit later.

3.3.3 Non-identical gates

Often channels are controlled by more than one protein, with each protein controlling a set of identical gates, but with the gates controlled by each protein different and independent. Consider, for example, the case of a channel with two types of gate, m and h, each of which may be open or closed. For the purposes of illustration, we will assume that the channel has two m subunits and one h subunit. Then, we let S_{ij} denote the proportion of channels with i of the of the m-gates open and j of the h-gates open (where $i = 0, 1, 2$ and $j = 0, 1$). Note, there are thus six possible configurations of the channel.

If m and h denote the proportions of each of the two types of gate that are open, then the law of mass action equations can be shown to be satisfied by

$$S_{00} = (1-m)^2(1-h), \quad S_{10} = 2m(1-m)(1-h), \quad S_{20} = m^2(1-h),$$

$$S_{01} = (1-m)^2 h, \quad S_{11} = 2m(1-m)h, \quad S_{21} = m^2 h,$$

so that the proportion of open channels is m^2h where m and h satisfy

$$\frac{dm}{dt} = \alpha(V)(1-m) - \beta(V)m, \qquad \frac{dh}{dt} = \gamma(V)(1-h) - \delta(V)h.$$

Note that here γ and δ are the rates of switching between the closed and open states for the h-gates (analogous to α and β for the m-gates).

We now have the required background to consider a model of nerve signal propagation.

3.4 The Hodgkin-Huxley model

The Hodgkin-Huxley model was developed by British scientists Alan Lloyd Hodgkin and Andrew Huxley in 1952 to explain the mechanisms underlying the initiation and propagation of action potentials (electrical signals) in the squid giant axon (the part of the squid's nervous system that controls its water jet propulsion system). It is considered one of the major achievements of mathematical biology, and Hodgkin and Huxley were awarded the Nobel prize in 1963 for their work.

More generally, the model can be used to describe the behaviour of a variety of excitable cells, including neurons, the cells which compose the electrochemical communication system that constituents our nervous system.

3.4.1 Structure of a neuron

Inputs detected by the dendrites are conducted to the soma. A nerve signal is then initiated at the axon hillock, which travels down the axon to terminal branches where the signal is passed to the next cells in the network. These signals are transmitted by action potentials, and cells which can transmit action potentials are called excitable cells (*e.g.* cardiac cells, smooth and skeletal muscle, secretory cells and neurons).

Neuronal signals travel along the cell membrane of the axon in the form of a local voltage difference across the cell membrane. In the inactivated state the cytoplasm (fluid inside the cell) in the axon is slightly negative in potential compared to the outside (-50mV difference) *i.e.* the cell is polarised, due to a difference in ionic composition, maintained by actively pumping sodium ions out of the cell and potassium ions into the cell, as well as differences in other ionic concentrations across the membrane.

It is tempting to think of neurons as a long electrical cable, but this is not quite right. The current in a neuron is made up of an ionic (rather than electron) flow, and the flow is across the membrane - i.e. transverse not longitudinal.

When the cell becomes partially depolarised a series of events takes place:

1. **The upstroke phase.** Sodium channels open in response to the depolarisation, allowing positively charged sodium ions to enter the cell, increasing the depolarisation further, till the cell becomes positively charged.
2. **The excited phase.** Over a slower timescale, the potassium channels open, allowing potassium ions to leave. Sodium ions continue to enter the cells and the potential difference slowly falls.
3. **The downstroke phase** The potassium ions make the cell negatively charged, which closes the sodium channels making the cell more negatively charged. The cell becomes hyperpolarised - it has overshot.
4. **The refractory and recovery phases** The sodium channels are now mostly inactive so cannot respond to any further stimulus. They gradually become active again, and the cell returns to its original state.

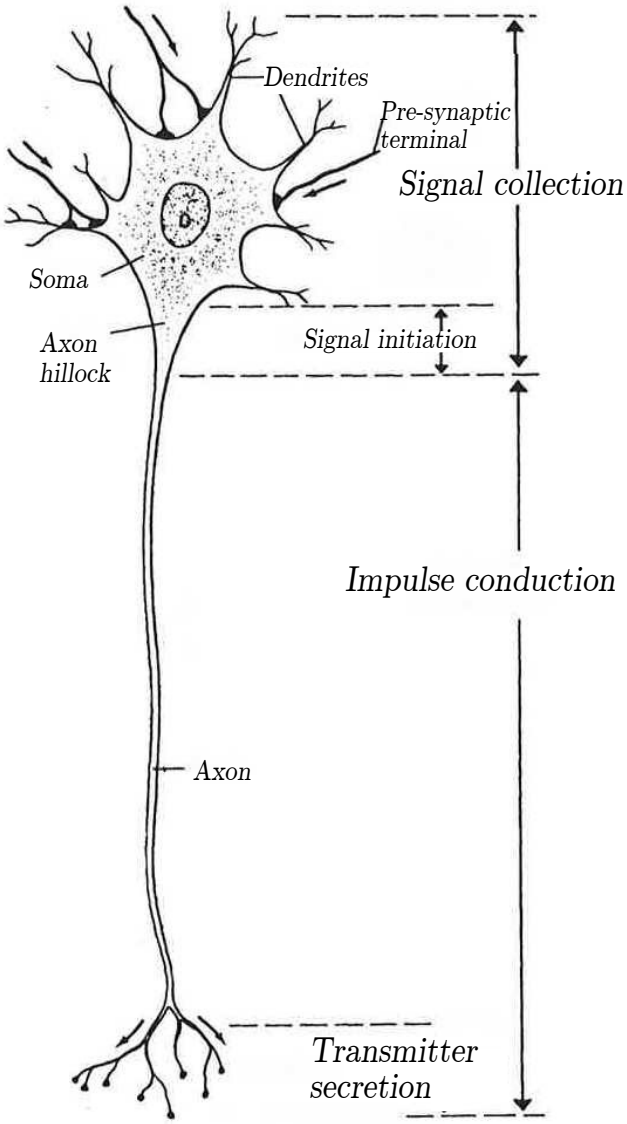

Figure 7: The structure of a neuron (reproduced from [2]).

5. **Propagation.** The nerve impulse propagates down the axon via the same process, since the sodium ions move down the axon along the potential gradient after entering the cell, depolarising the next bit of the cell and hence triggering the same process. We have a **travelling wave**.

3.4.2 Hodgkin-Huxley model for a clamped neuron

We make the simplifying assumption that the axon is **clamped**. This ensures there are not spatial variations of the variables of interest. An axon can then be represented schematically as:

where we remember that the axon is cylindrical. We have condutances to flow g_K, g_{Na} and a capacitance, C due to the thickness of the membrane. We define:

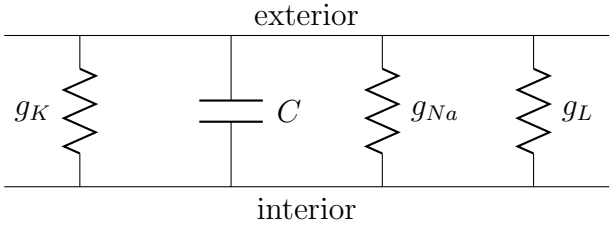

Figure 8: Schematic representation of the axon

- q - charge density;
- C - capacitance per unit area;
- a - radius of axon;
- I_i - rate of movement of ions from exterior to interior per unit membrane area;
- V - departure from resting voltage.

Then $q = 2\pi aCV$ and hence

$$\frac{dq}{dt} = -2\pi a I_i = -2\pi a \left(\underbrace{I_{Na}}_{\text{sodium current}} + \underbrace{I_K}_{\text{potassium current}} + \underbrace{I_L}_{\text{leakage current (other ions)}} \right),$$

where:

$$I_{Na} = g_{Na}(V - V_{Na}), \qquad I_K = g_K(V - V_K), \qquad I_L = g_L(V - V_L).$$

Thus, on substituting we obtain

$$\frac{dV}{dt} = -\frac{1}{C}\left(g_{Na}(V)(V - V_{Na}) + g_K(V)(V - V_K) + g_L(V - V_L)\right),$$

where we have assumed the conductances of sodium and potassium ions to depend on the current voltage whilst the conductance of other ions is constant. This gives the first equation of the Hodgkin-Huxley model. It remains to specify g_{Na} and g_K.

We introduce the sodium activation variable, m, the sodium inactivation variable, h, and the potassium activation, n, and write:

$$\tau_m(V)\frac{dm}{dt} = m_\infty(V) - m,$$

$$\tau_h(V)\frac{dh}{dt} = h_\infty(V) - h,$$

$$\tau_n(V)\frac{dn}{dt} = n_\infty(V) - n,$$

so for constant V, $m \to m_\infty(V)$ exponentially with time constant, τ_m, *etc.* . Since m, n are activations and h is an inactivation we choose:

and take

$$g_{Na} = \bar{g}_{Na} m^3 h, \qquad g_K = \bar{g}_K n^4.$$

These forms for g_{Na} and g_K are chosen to fit the data.

The system consists of four highly nonlinear coupled ODEs and so is very difficult to understand / analyse (especially considering digital computers where still in their infancy in the 1950s). We shall follow Fitzhugh and Nagumo and apply phase-plane methods to a simplified version of the system.

3.4.3 The Fitzhugh-Nagumo analysis of the Hodgkin-Huxley model

We now reduce our four-equation model to a pair of coupled equations.

We can write

$$C\frac{dV}{dt} = -\left(\bar{g}_{Na} m^3 h + \bar{g}_K n^4 + \bar{g}_L\right)(V - V_{eq}),$$

where

$$V_{eq} = \frac{\bar{g}_{Na} m^3 h V_{Na} + \bar{g}_K n^4 V_K + \bar{g}_L V_L}{\bar{g}_{Na} m^3 h + \bar{g}_K n^4 + \bar{g}_L},$$

so the time constant for V is

$$\tau_V = \frac{C}{\bar{g}_{Na} m^3 h + \bar{g}_K n^4 + \bar{g}_L} \sim 1\,\text{ms}.$$

Now, we have

$$\tau_h \sim \tau_n \gg \tau_V \gg \tau_m.$$

As τ_m is small we take $m \approx m_\infty(V)$ and as $\tau_h \sim \tau_n$ and $n_\infty + h_\infty \approx const = \bar{h}$ we have

$$\tau_n \frac{d}{dt}(n + h) = \bar{h} - (n + h) \quad \Rightarrow \quad n + h = \bar{h}.$$

The system is then

$$C\frac{dV}{dt} = -\left(\bar{g}_{Na} m^3 (\bar{h} - n)(V - V_{Na}) + \bar{g}_K n^4 (V - V_K) + \bar{g}_L (V - V_L)\right),$$

$$\tau_n(V)\frac{\partial n}{\partial t} = n_\infty(V) - n.$$

We shall take τ_n to be constant, and nondimensionalise

$$V = V_{Na}\tilde{V}, \qquad t = \tau_n \tilde{t},$$

where tildes indicate dimensionless quantities (and we note n is already dimensionless). The dimensionless system is then (dropping tildes)

$$\frac{\partial n}{\partial t} = n_\infty - n.$$

$$\epsilon \frac{dV}{dt} = -g(V, n),$$

where $epsilon = C/\bar{g}_{Na}\tau_n$,

$$g(V, n) = \gamma_K(V + V_K^*)n^4 + \gamma_L(V - V_L^*) - (1_V)(\bar{h} - n)m^3,$$

and $\gamma_K = \bar{g}_K/\bar{g}_{Na}$, $\gamma_L = \bar{g}_L/\bar{g}_{Na}$, $V_K^* = -V_K/V_{Na}$ and $V_K^* = V_L/V_{Na}$.

We will neglect γ_L except when V is close to zero, as it is small.

3.4.4 Phase plane analysis

We now investigate the behaviour of the simplified system by considering the phase plane - *i.e.* the (V, n) plane.

If $\epsilon \ll 1$ then $g \to 0$ very quickly. We plot the nullclines, which are given by

$$\frac{dn}{dt} = 0, \quad \Rightarrow \quad n = n_\infty(V),$$

$$\frac{dV}{dt} = 0, \quad \Rightarrow \quad g(V, n) = 0.$$

We hence need to sketch the curves

$$g = 0 \quad \Rightarrow \quad \frac{n^4}{\bar{h} - n} = \frac{(1 - V)m^3(V)}{\gamma_K V_K^*(1 + \frac{V}{V_K^*}},$$

where we have neglected γ_L.

We have one fixed point

$$V^* = 0, \qquad n^* = n_\infty(0).$$

The Jacobian matrix is then

$$J = \begin{pmatrix} -1 & n_{\infty_V} \\ -g_n & -g_V \end{pmatrix}$$

where the subscripts indicate a partial derivative.

Then

$$\operatorname{tr} J = -1 - g_V, \qquad \det J = g_V + n_{\infty_V} g_n.$$

where $n_{\infty_V} > 0$ and

$$g_V = \gamma_K n^4 + \gamma_L + (\bar{h} - n)m^3 > 0, \qquad g_n = 4\gamma_K V_K^* n^3 + (1 - V)m^3 > 0$$

if V is sufficiently close to zero.

Hence $\operatorname{tr} J < 0$, $\det J > 0$ and so we have a stable fixed point. Thus a sufficiently large perturbation to V excites an action potential.

3.4.5 Generalised Fitzhugh-Nagumo

Consider the system

$$\epsilon \frac{dV}{dt} = I^* + f(V) - w,$$
$$\frac{dw}{dt} = \gamma V - w,$$

where $f(V) = V(V-a)(1-V)$ since we need f to be cubic for the system to be excitable. This comes from the reduction of the Hodgkin-Huxley model with $g = w - f(V)$, $n_\infty = gammaV$, $w = n$.

When $I^* = 0$ the nullclines satisfy

$$w = V(V-a)(1-V), \qquad w = \gamma V.$$

Hence we have one fixed point, and an action potential.

If, however, we have $I^* \neq 0$ we have three possibilities:

and if γ is not large enough, then we have three fixed points.

We can see that we can get different behaviour depending on the shapes of the graphs (changing I^* gives rise to bifurcations), but we have successfully explained the existence of action potentials.

References

[1] NJL Brown, AD Sokal, and HL Friedman. The complex dynamics of wishful thinking: The critical positivity ratio. *Am. Psychologist*, 68(9):801813, 2013.

[2] L. Edelstein-Keshet. *Mathematical models in biology.* SIAM, 2005.

[3] JD Murray. *Mathematical Biology: I. An Introduction.* Springer, 2003.

[4] L. A. Segel. On the validity of the steady state assumption of enzyme kinetics. *Bulletin of Mathematical Biology*, 50:579–593, 1988.

[5] G. I. Taylor. The formation of a blast wave by a very intense explosion. ii. the atomic explosion of 1945. *Proc. R. Soc. (Lond.) A*, 201:175–186, 1950.

[6] D. W. Thompson. *On Growth and Form.* Cambridge University Press, 2nd edition, 1942.