Modeling with Dynamical Systems and Kinetic Equations

Reinhard Illner*
Department of Mathematics and Statistics
University of Victoria
Victoria, BC V8W 3P4

May 28, 2007

Abstract

This article contains both a survey of and some novelties about mathematical modeling problems which emerged within recent years in physical chemistry, microbiology, and multi-lane traffic flow. Specifically, we first present a generalization of the Kolmogorov-Avrami model for crystallization dynamics for cases where the crystallization is incomplete and the classical model fails; second, the concept and an application of transcriptional-translational oscillators operating inside a living cell. The feasibility and significance of such oscillators is an important topic in molecular biology, and, as will be shown, their interactions may be the cause underlying phenomena like circadian rhythms. The basic equations are stochastic differential equations including Markovian random variables, and their associated Kolmogorov master equations are a linear kinetic system of PDEs. Third, we engage in a discussion of traffic flow models for the multi-lane scenario, with emphasis on Fokker-Planck type systems and their properties. We summarize and discuss results from a series of papers in which such models were introduced as alternatives to other, rather different kinetic models, and we conduct a comparison. There are discussions on the relationship between kinetic and macroscopic models, on fundamental diagrams, and on modeling ingredients which may lead to more than one equilibrium solution, a scenario known as a bifurcated

*Supported by a grant from the Natural Sciences and Engineering Research Council of Canada

(or multi-valued) fundamental diagram. We go though an extensive presentation of the properties and degeneracies of the new models, and we briefly introduce relative entropy and discuss its applications.

1 Introduction

When I was first asked to lecture on kinetic theory in Porto Ercole I hesitated before accepting, not because I did not want to participate, but because I was not sure that what I have to contribute would be quite appropriate. After all, my last real contributions to the theory of the Boltzmann equation are 10 years old and somewhat dated; I spent several years after that exploring a generalization of the Vlasov-Poisson system which we called the Vlasov-Manev system (there is a $1/r^2$ correction to the Newtonian potential in this system); this system leads to a lot of serious mathematics but is dubious from the point of view of physical applicability. For that reason the Vlasov-Manev system has not found much attention in the community.

When I finally agreed to be one of the lecturers, it was under a different pretext, namely that of mathematical modeling with motivating applications from the outset. This has been a guiding motive for my own research over the last few years, and it has been a consequence of an "open door" policy that I am known for– namely, colleagues who explore a scientific problem that may involve a dynamical system or kinetic model are welcome to knock at my door, and I will see what can be done about the problem. This has led me astray from my traditional research but brought me in touch with most interesting scientific projects, and I would like to use this article to present a survey over three of the most rewarding ones. I will have to skip technicalities and many other details, but the interested reader may find most of these in the available publications on the various models.

As announced in the abstract, the three themes are incomplete crystallization, genetic (transcriptional-translational) oscillators and their possible biological consequences, and kinetic models for multi-lane traffic flow. Each subject will be introduced at length in the relevant section, so I will be brief here. However, there are some novelties which I observed as I wrote this. In Section 2, the treatise on incomplete crystallization, I noticed that the case of more general stoichiometry was not adequately covered in the previous publications, so it is done here. This problem is in some sense a "warmup" and really includes neither dynamical systems theory nor kinetic theory; rather, it is an exercise in probabilistic modeling, at a fairly elementary level, but with satisfying results.

The treatise on the Kolmogorov master equations for a simple example

(with the same structure as transcriptional-translational oscillators) in Section 3.7 is cursory and incomplete. While the process is rather standard, it appears to be novel in this context. The case where the transition rates are state-dependent resembles exactly the scenario where a kinetic system contains fast and slow time scales, and where the fast time scales average out in a fluid dynamic approximation. The details in the case of the TTO model are a topic of current research.

The section of multi-lane traffic flow contains a literature survey and a detailed discussion of Fokker-Planck type kinetic models. Some recent results are announced, in particular the newly discovered link between this kind of kinetic model and the Aw-Rascle macroscopic model. Studies on traffic flow open a very open-ended field of research, and they are popular because *everybody is an expert.*

There is probably little really new mathematics in any of these sections; the novelty is in the application rather than the tool. Of course, Applied Mathematics should never be an end in itself but almost always a means to an end, usually in science and engineering. The ends here are in physical chemistry, in microbiology and in traffic engineering.

2 Incomplete Crystallization

In this first lecture we present the solution to an unusual example of crystallization dynamics first observed at the University of Victoria ([14],[15]) about 10 years ago. The problem involved instability of an *unstable crystalline binary phase* of $CO_2.C_2H_2$, which formed at 90 degrees Kelvin when a mixture of carbon dioxide and acetylene was sprayed onto a glass panel. Spectroscopy showed that a new type of crystal (the binary phase) formed, but did not last; rather, over a period of about 6 hours pure crystalline CO_2 formed, embedded in an amorphous matrix of C_2H_2. Furthermore, this transition was not well modeled by the classical Kolmogorov- Avrami crystallization curve; eventually, the chemists who had conducted the experiments contacted the mathematics department to see whether we could develop better models. We could, and we will show in this first lecture what was done. We begin by reviewing the classical theory.

2.1 Complete Crystallization: The Kolmogorov-Avrami Model

The Kolmogorov- Avrami model produces a curve which predicts what fraction of a crystallizable substance will have crystallized by time t. This may

be found in textbooks on physical chemistry. The model is also very satisfactory in the sense that the curves match data very well in cases where the crystallization is complete.

It is quite easy to show a derivation of the model, and we do this for completeness. Crystallization (of, say, liquid CO_2) will start at certain sites which we will refer to as crystallization nuclei (or "impurities"). In the absence of such nuclei the substance will stay a liquid. From each nucleus an individual crystal, which we will call a globule, will grow outward at constant radial crystallization speed $v > 0$; eventually these globules will impinge upon each other, thereby stopping the growth in each others direction. This process will continue until there is no liquid left. Figure 1 shows a snapshot of a two-dimensional simulation.

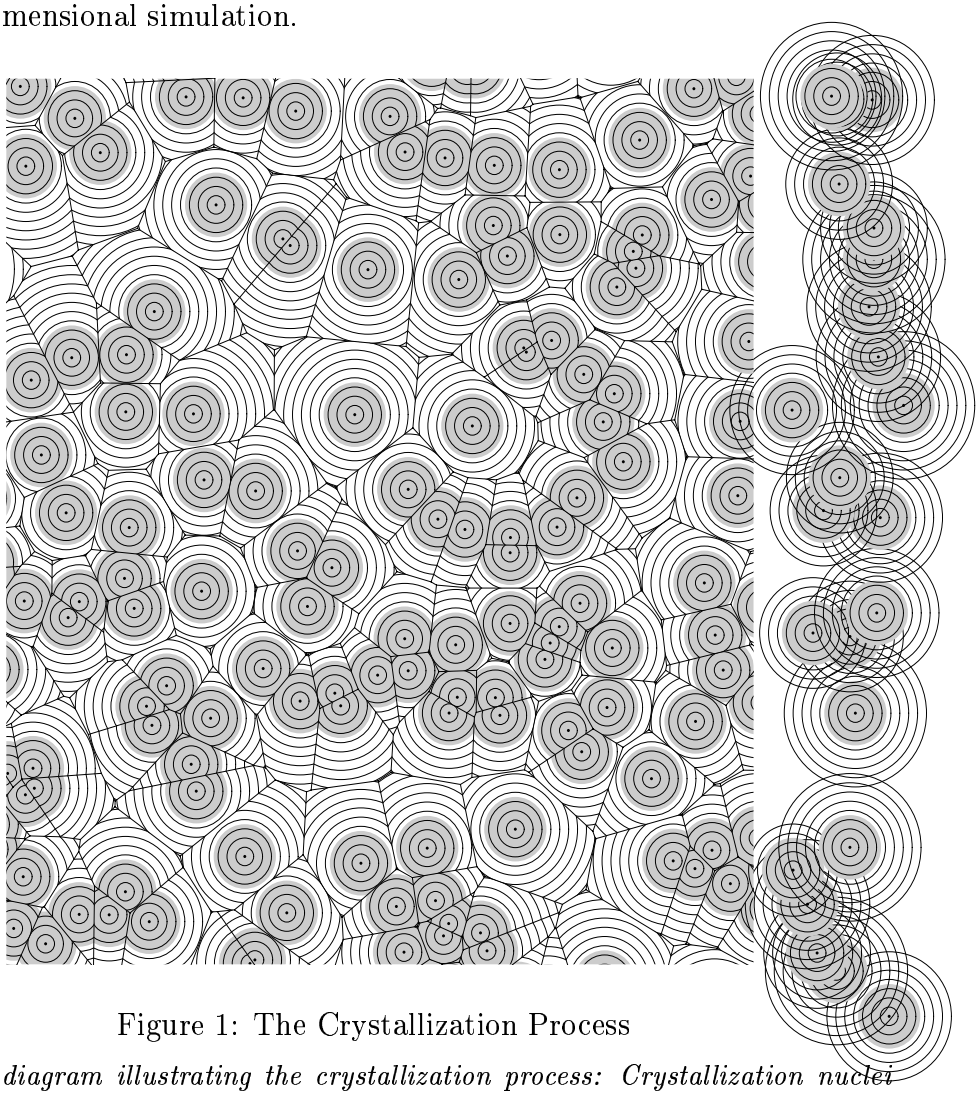

Figure 1: The Crystallization Process

A Voronoi diagram illustrating the crystallization process: Crystallization nuclei grow until they impinge upon one another.

We begin by assuming that N nuclei are independently equidistributed in a (large) domain D with volume V. Let $Q_1, \ldots, Q_N$ denote the locations of the nuclei, and let P be an arbitrary reference point in D. If we set $X = \min_{i=1\ldots N} \mathrm{dist}(P, Q_i)$, then elementary probability theory shows that for $a > 0$

$$Prob\{X > a\} = \left(1 - \frac{4\pi a^3}{3V}\right)^N.$$

We can then immediately write down the cumulative distribution function of X, defined by $F_N(a) = Prob\{X \leq a\}$, and find

$$F_N(a) = 1 - \left(1 - \frac{4\pi a^3}{3V}\right)^N.$$

In the limit $N \to \infty$, $V \to \infty$ such that $N/V = \lambda > 0$ we obtain a spatial Poisson process with intensity $\lambda > 0$, and rewriting $F_N(a) = 1 - \left(1 - \frac{4\pi\lambda a^3}{3N}\right)^N$ we see that in this Poisson limit we find $F(a) = lim_{N\to\infty} F_N(a)$ as

$$F(a) = 1 - e^{-4\pi\lambda a^3/3}.$$

Returning to the crystallization question, we see that the reference point P will have crystallized by time t exactly if its distance from the closest nucleus is smaller than vt. This will have happened with probability $F(vt)$, and the crystallization probability for P by time t is therefore

$$\varphi(t) = 1 - e^{-kt^3} \tag{1}$$

where $k = 4\pi\lambda v^3/3$. If we average over all equidistributed points P in D Equation (1) does not change, and gives therefore the expected fraction of volume which will have crystallized by time t. Equation (1) is known as the Kolmogorov-Avrami model.

Remarks.

- Usually, neither λ nor v are known a priori. Notice that small λ means a slower crystallization rate.
- Our derivation applies to three dimensions. The corresponding formula in two dimensions is $\varphi(t) = 1 - e^{-kt^2}$.
- In practice, chemists sometimes uses variations of the model with higher powers of t, for example $\varphi(t) = 1 - e^{-kt^{3.2}}$. Reasons given for this are that the growth around some nuclei may start with a delay, or that

(equivalently) nuclei may form spontaneously while the crystallization process is already under way. Moreover, while our derivation tacitly assumes that the globule growth is radial, this is in general not the case; the microcrystals may grow in simple or complicated geometric shapes, and new nuclei may form at the edges of these shapes. The variations used by chemists seem to address such complications reasonably well.

2.2 An Experiment where the Kolmogorov-Avrami Model failed

An unusual example of crystallization dynamics was observed at the University of Victoria ([14], [15]) about 10 years ago. T. Gough and collaborators discovered a hitherto unknown *crystalline binary phase* $CO_2.C_2H_2$, formed at 90 degrees Kelvin when a mixture of carbon dioxide and acetylene was sprayed onto a glass panel. Spectroscopy showed that a new type of crystal (the binary phase) formed, but did not last; rather, over a period of about 6 hours pure crystalline CO_2 formed, embedded in an amorphous matrix of C_2H_2. Furthermore, this transition was not well modeled by the classical Kolmogorov- Avrami crystallization derived in the previous section.

They tested their data for compatibility with the Kolmogorov-Avrami model by using log-log plots. Specifically, note that if $\varphi(t) = 1 - e^{-kt^n}$, then, by setting $t = e^s$, and taking double logarithms, it follows that

$$\ln(-\ln(1 - \varphi(e^s))) = \ln k + ns.$$

Setting $C = \ln k$ and denoting the left-hand side by $f(s)$, the identity simplifies to

$$f(s) = C + ns,$$

i.e., in this representation the crystallization curve becomes a straight line, and, if the model applies, the properly rescaled data should also fall approximately on a straight line. However, this was not at all the case for the data from the decomposition of $CO_2.C_2H_2$, and it was clear that the classical theory was insufficient to explain the time dynamics of this phenomenon.

2.3 Generalized Models

The derivation of the Kolmogorov-Avrami model given earlier was based only on the probabilities that an arbitrary reference point P would be at a certain distance from the closest nucleus. This works very well for situations where the whole substance will crystallize, but such is no longer the case for the

example presented in the previous section. We have to be more careful in the model design.

A moment's thought shows that the Voronoi diagam partition of the domain D defined by the nuclei Q_i should be of crucial importance for the process. Recall that the Voronoi cell D_i associated with Q_i is the set of all points R such that $|R-Q_i| < min_{j\neq i}|R-Q_j|$, or, in words, R is closer to Q_i than to any other nucleus. The partition of D obtained in this way is called the Voronoi diagram associated with the nuclei.

For the time being we will focus on the growth of one crystalline globule which starts at Q_i. This growth rate will initially be cubic, but as soon as the globule reaches a boundary of the Voronoi cell it will stop its growth in this direction, because by definition of the Voronoi boundary the growth from a neighboring nucleus will reach that boundary simultaneously, and the globules will impinge upon each other. See Figure 1.

The stoichiometry of our original compound (in the example, the binary phase $CO_2.C_2H_2$) is of importance for the sequel. Suppose that fraction $q < 1$ of the available volume can actually crystallize, while the remaining fraction $1-q$ will be occupied by the amorphous "waste" component. In the example we have $q = 1/2$. Considering still one Voronoi cell, we now make an idealized ansatz for the individual growth curve of the globule: Let V_c be the volume of the Voronoi cell, set $s := qV_c, b := (s/k)^{(1/3)}$ and define

$$g(s,t) = \begin{cases} kt^3 & \text{for} \quad t \leq b \\ s & \text{for} \quad t > b \end{cases} \tag{2}$$

The constant k is proportional to the cube of the radial growth speed of the globule. This is a growth curve which "pretends" that cubic growth continues until the crystallizable fraction of the cell is full; at that point the growth stops completely. In reality, the growth will not be cubic after one or more boundaries of the Voronoi cell are reached; at this point growth will slow down and stop completely at $t = b$, when the maximal possible fraction of the cell has crystallized. The details of the true individual growth curve will depend on the geometry of the Voronoi cell in question, and on q. The idealized growth curve is more realistic for small q because the bulk of the Voronoi cell will be filled with the amorphous residue, and the crystallization will stop rapidly. If q is large, matters are more complicated, as the globule may impinge into many Voronoi boundaries before the crystallization stops.

The key idea of a generalized model is to average idealized individual growth curves as in (2) over the volume distribution of Voronoi cells generated by a Poisson process. Specifically, if we assume as before that the crystallization nuclei are distributed according to a Poisson process with intensity $\lambda > 0$, the *crystallizable* fractions of the Voronoi cells will have volu-

mina which are distributed according to some (unknown) probability density $f_q(s)$, and the averaged crystallization curve will be an integral

$$\varphi(t) = \int_0^\infty g(s,t) f_q(s) ds. \tag{3}$$

Note that this growth curve is not dimensionless–by construction, it has the dimension of volume. We can divide by a normalizing constant to rewrite φ as a fraction of the total crystallizable volume. Growth curves of the type (3) have the potential to model most incomplete crystallization curves well. However, the following problems arise and must be addressed:

- the idealized growth curves (2) are not very realistic; specifically, they are expected to be less and less realistic as q approaches 1.
- the distribution densities f_q are not known. In fact, it is a hard open problem in computational geometry to establish good approximations of statistical properties of Voronoi cells, including their volumina. A reasonable approximation for f_q for the case $q = 1$ is a density function

  $$f(s) = \beta s^2 e^{-\gamma s^2},$$

  where β and γ are linked by the constraint that $\int_0^\infty f \; ds = 1$ (see Ref. [18]).

If the idealized growth curves (2) are used in (3), an easy calculation shows that

$$\varphi(t) = \int_0^{kt^3} s f_q(s) \; ds + kt^3 \int_{kt^3}^\infty f_q(s) \; ds \tag{4}$$

For the example of the disintegration of $CO_2.C_2H_2$ into crystalline CO_2 and residual C_2H_2 the stoichiometry is 1-1, so $q = 1/2$. in Refs. [14],[15] we used $f(s) = \beta s^2 e^{-\gamma s^2}$ in (4) to produce a theoretical growth curve, and then scaled time and β to produce a best fit to the data. The result was completey convincing: See Figure 2.

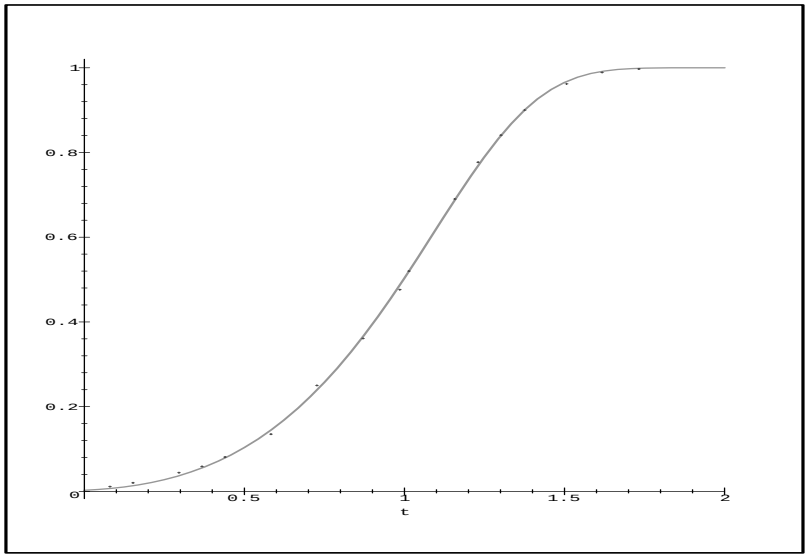

Figure 2: The growth curve matched to data points

Remarks.

- Since this research was first published, many more examples for incomplete crystallization as described here have been discovered by chemists, and the growth curves of type (4) provide good matches for their growth curves (T. Gough, personal communication).

- If one chooses for f_q the density distribution of the volume of the largest sphere centered at the nucleus which will fit into a Voronoi cell, Formula (4) reproduces the Kolmogorov-Avrami model. This is not surprising, because this approach essentially ignores the space available to the waste product. It is not hard to see that this largest sphere will on average fill only 1/8 of the volume of the cell. See [18] for more details.

- The topic we discussed in this section doesn't really fit the label "kinetic theory." We deemed it of sufficient interest to be presented here; there is of course the central aspect of "averaging," in this case over all the Voronoi cells in the volume at hand.

3 Transcriptional-Translational Oscillators, Ultradian, and Circadian Rhythms

Our second subject brings us from physical chemistry into the domain of microbiology, specifically cell biology. The section will provide an introduction to coupled gene expression processes which provide oscillations in protein or protein complex concentrations; such oscillations are observed in nature in both eukaryotes and prokaryotes. Their periods are typically of the order of magnitude of 15 minutes to a few hours, significantly shorter than the 24 hour day/night cycle, and hence they are known as "ultradian" oscillations.

We will present a mathematical model based on realistic biochemical processes occuring in the cell, thus *possibly* a true mechanism; the complexity of cell biology is such that experimental verification is very difficult. Dozens, if not hundreds, of the biochemical processes we use in the model occur in real cells simultaneously at any time.

3.1 The Setting

This article is an article on mathematical modeling, not microbiology, so we find it appropriate to include a few basic (and trivial, yet impressive) facts about cell biology. First, recall that each individual cell contains the genome (DNA) molecule consisting of two DNA strands that in turn encode all the genetic information about the living being. For a human, the length of a single (untangled) DNA molecule would be on the order of magnitude of 2 meters; given that the human body contains about 10^{13} cells, the accumulated DNA molecules of one human being, laid end to end, would reach from the Earth to the Sun and back 70 times. It would only take the DNA of a little over 2,000 humans to reach to the next solar system.

Of course, the DNA molecule is tangled (knotted) in unimaginable complexity to fit inside the cell, where it controls and drives life processes. The crucial processors are the genes, which should be thought of as short pieces of ("sites" on) the DNA strand. These sites drive chemical processes leading to the production of proteins and protein complexes, a procedure known as "gene expression."

In the sequel we describe how the expression of two genes via messenger RNA, translation into proteins, and protein reactions to form secondary protein compounds, which interact with "the other gene" on the DNA strand, can produce ultradian oscillations. We call such an oscillating system a primary oscillator or transcriptional-translational oscillator, or in short a TTO. Later we will describe how couplings between TTOs can lead to secondary os-

cillations with circadian periods; this is a possible explanation for the cellular mechanisms causing circadian rhythms.

A gene may be active or passive; if it is active, it will produce messenger RNA (mRNA) at a steady rate, a process called transcription. The gene state (active or passive) typically depends on whether a site-specific protein complex (a protein "dimer") is attached to the site (we call the site "occupied"). If the presence of the specific dimer causes transcription, we say the dimer "activates" the site. The opposite is also possible – occupation by a site-specific dimer may stop transcription. In this case we speak of site repression or inhibition. The presence of an activating or inhibiting dimer at each site should be modeled as a Markov process whose parameters depend on the concentrations of the dimers in the cell (actually, it is more complicated; we will discuss this further, after the TTO modelling is complete). Two genes may "communicate" if the dimers produced by their transcriptional-translational reactions are site-specific for the other gene. This is the basic idea behind a TTO.

The job of the mRNA molecules transcribed by the gene is to assemble raw materials available in the cell into proteins. This process is called translation, and the chain transcription-translation is known as gene expression. The proteins may react with proteins of the same kind to form homodimers, or with different proteins to form heterodimers. It is these dimers (protein complexes) which may reattach to the DNA strand, then travel along it until they arrive at one of their specific sites and assume their task of activation of repression.

In our model of a TTO, two genes (DNA sites) are transcribed into mRNA, and this process is a starting point of the following cyclical chemical dynamics.

- Transcription by gene 1 occurs when site 1 (its regulatory region) is unoccupied. Its state is given by a random variable X_1, so that

  $X_1 = 0$ if site 1 is empty; $X_1 = 1$ if site 1 is occupied by D_2 (see below)

- When gene 1 is active it produces mRNA (measured in molecules per cell, R_1) at a constant rate k_{11}. These molecules undergo first-order decay with a rate constant k_{12}.

- The mRNA molecules are translated with rate constant k_{13} into protein P_1, which: (a) decays at rate constant k_{14}, (b) forms homodimers D_1 at rate k_{16}, and (c) forms heterodimers D_{13} with proteins P_3 from a third gene (see below) with a rate constant k_{61}.

- The homodimer D_1 binds to site 2, and thereby activates the transcription of gene 2. The state of gene 2 is given by the value of a random variable Y_1 so that

$$Y_1 = 0 \text{ if site 2 is empty, and } Y_1 = 1 \text{ if site 2 is occupied by } D_1.$$

- When activated, gene 2 transcribes its mRNA; its mRNA translates into protein P_2, which forms homodimer D_2, which in turn feeds back to inhibit gene 1 (above). In addition, the P_2 molecules decay with a certain (biological) half-life.

- These linked reactions generate a TTO for an appropriate choice of parameters. The parameters used in our subsequent calculations are listed in Table 1. Our model entails gene 1 being inhibited by homodimer D_2 and gene 2 being activated by homodimer D_1. This is the mechanism leading to *primary* oscillations.

Figure 3 is a caricature of the process.

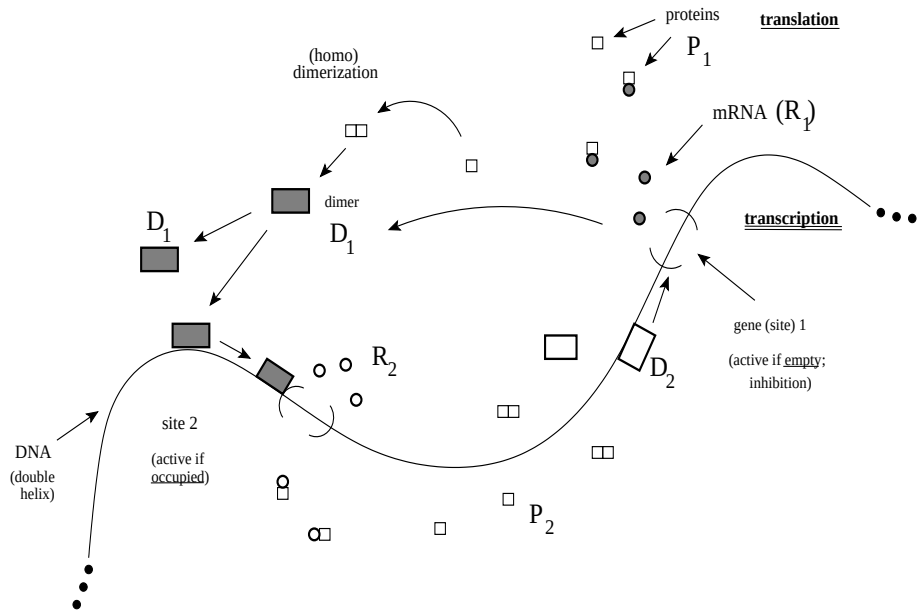

Figure 3: A caricature of a TTO

3.2 Stochastic and Time-averaged ODE Systems for Primary Oscillators

We denote by $R_i, P_i, D_i, \quad i = 1, 2$ the numbers of mRNA, translated protein and homodimer molecules per cell produced by site i. The above scenario is then summarized in the following system of stochastic differential equations (only two of the equations contain the random variables X_1 and Y_1 explicitly, but all dependent variables are then of necessity random variables). The parameters k_{11} etc. are listed in Table 1. The notations for these parameters are changed from those of Ref. [8] for a more logical sequence. For a complete list of biologically reasonable parameter values, see the reference.

$$
\begin{aligned}
R_1' &= k_{11}(1 - X_1) - k_{12}R_1 && (5)\\
P_1' &= k_{13}R_1 - k_{14}P_1 - 2k_{15}P_1^2 + 2k_{16}D_1 - k_{61}P_1P_3 + k_{62}D_{13} && (6)\\
D_1' &= k_{15}P_1^2 - k_{16}D_1 && (7)\\
R_2' &= k_{11}Y_1 - k_{12}R_2 && (8)\\
P_2' &= k_{23}R_2 - k_{24}P_2 - 2k_{25}P_2^2 + 2k_{26}D_2 && (9)\\
D_2' &= k_{25}P_2^2 - k_{26}D_2 && (10)
\end{aligned}
$$

The loss terms in the various equations reflect the fact that mRNA, protein and dimer all have a biological half-life in the cell. For example, the half-life of the homodimer D_1 is $T = \ln 2/k_{16}$. The first term in the second equation describes the translation process; note that there is no reason why k_{12} and k_{13} should be coupled; the former, as mentioned, is the rate of disintegration of an mRNA molecule, while the latter is the rate at which such a molecule manufactures proteins. The last two terms in the second equation reflect the combination of proteins P_1 and P_3 (which is produced by the second primary oscillator) to form the heterodimer D_{13}. This heterodimer in turn breaks down into pairs P_1 and P_3 at rate constant k_{62}.

- Notice that there is no coupling between the equations for R_1, P_1, D_1 and R_2, P_2, D_2 except possibly through the laws of the random variables X_1, Y_1. This is exactly how the coupling is achieved, and we will shortly explain how.

- The dimensions we choose for our model variables are in molecules/cell.

A second primary oscillator will be given by a nearly identical set of equations, except that the periods of the oscillations are slightly different. This can, of course, be achieved by changing the parameters in many ways, but

Parameter	Value	Dimension
k_{11}	1800	$1/hr$ (1 mRNA every 2 seconds)
k_{12}	3.2	$1/hr$ (half-life $T \approx$ 13minutes)
k_{13}	700	$1/hr \times$ molecule
k_{14}	4	$1/hr$
k_{15}	3.6×10^{-4}	$1/(hr \times \text{molecule}^2)$
k_{16}	15	
$k_{23} = k_{13}$		
k_{24}	4	
k_{23}	1400	
k_{25}	10^{-4}	
k_{26}	5	

Table 1: Parameters. The dimensions are $[k_{12}] = hr^{-1}, [k_{12}] = (nr. \times hr)^{-1}$, $[k_{15}] = (nr.^2 \times hr)^{-1}$

the simplest method is to have the two TTOs identical in nature but with different time scales. To do this we multiply each right hand side by a fixed constant $\delta > 0$, where δ is close (but not identical) to one. For example, the first equation of the second oscillator will read

$$R_3' = \delta(k_{11}(1 - X_2) - k_{12}R_3).$$

The parameters chosen reflect, where available, reasonable choices of known molecular processes. The critical ones for establishing the periods of the primary oscillators turn out to be the decay times of the mRNAs and proteins (not surprisingly, the oscillations are based on a Hopf bifurcation, which is driven by the values of these parameters. We discuss this aspect later). For the former, a half-life of 13-17 minutes and for the latter, 4-17 minutes generate ultradian oscillations in the model. Some of the values used in the simulation are given in Table 1.

The coupling between the two sites communicating in each oscillator is, as stated, provided by the random variables X_i, Y_i. The times for which these random variables stay constant are assumed to be exponentially distributed and follow *"their own time scale."* For example,

$$Prob\{X_1 = 0 \quad \text{in } (t, t+h) | X_1(t) = 0\} = \exp(-D_2(t)h/\epsilon) + o(h),$$

$$Prob\{X_1 = 1 \quad \text{in } (t, t+h) | X_1(t) = 1\} = \exp(-rh/\epsilon) + o(h)$$

while

Parameter	Value
r	25
s	5000
q	5500
δ	1.125

Table 2: More parameters. We set $[\epsilon] = hr$, so that $r, s, \ldots$ become dimensionless

$$Prob\{Y_1 = 0 \quad \text{in } (t, t+h) | Y_1(t) = 0\} = \exp(-D_1(t)h/\epsilon) + o(h),$$

$$Prob\{Y_1 = 1 \quad \text{in } (t, t+h) | Y_1(t) = 1\} = \exp(-st/\epsilon) + o(h).$$

ϵ is a time scaling parameter, introduced for convenience to express the fact that the binding and unbinding of the homodimers occurs on a faster time scale than the remaining processes. The constants r and s measure, relative to the scale ϵ, the average times for which the sites will remain occupied. As this is an internal parameter of the site it should not depend on the states of the rest of the system (like, for example, the dimer concentrations). We give values for these parameters, and for δ, in Table 2.

The parameter ϵ scales the most critical parameter, namely the rate constant for binding of the transcriptional-regulatory proteins (D_1, D_2) to the sites of the genes. There is experimental work ([2], [29], [35]) which indicates macroscopic (observed) second-order binding rates of $10^{10}/(M \cdot \text{s})$. In terms of molecules per cell in a bacterium, this translates into approximately $36,000$ molecules/(cell$\cdot$hour) (1 molecule per bacterial cell is about $10^{-9}\ M$). This rate is much larger than predicted for a diffusion-limited reaction, and is believed to be due to the unusual mechanism of association, in which the protein binds with low specificity to unrelated DNA and then migrates along the DNA molecules to its high-affinity regulatory site by one-dimensional diffusion. It has been shown that this accounts for the rate enhancement observed in [2].

We therefore set the average "free" time of the binding site for D_2 as ϵ/D_2, and the average "occupied" time as ϵ/r. Their quotient is independent of ϵ, but will change with the homodimer concentration D_2.

As stated, the average times for which a dimer stays bound ($\epsilon/r, \epsilon/s$, etc.) are independent of the state of the system. In contrast, the "free" times are inversely proportional to the concentration of the attaching homodimer. In one of the simulations given in [8] we used $r = 25$ and $\epsilon = 10^{-1}$sec (which corresponds to $\frac{\epsilon}{r} = \frac{1}{250}$sec, or an average of 900,000 binding events per hour).

The corresponding stochastic simulation compares well with a (yet to be described) limiting scenario for which $\epsilon = 0$. Before we describe this limiting scenario in detail we present the remaining equations making up the complete oscillatory system.

The protein products P_1 and P_3 of the first and second primary oscillators combine to produce a heterodimer D_{13}. Assume that this heterodimer binds to the regulatory site of a fifth gene and activates it for transcription (other constructs, involving other heterodimeric products of the two primary oscillators, and either stimulation or inhibition of transcription of the fifth gene, could also be used). Transcription, translation, and dimerization of the protein product of gene 5 yields the product D_5, the primary circadian output of the model (although all variables show circadian behaviour to a greater or less extent, as seen in numerical experiments).

The corresponding system is

$$\begin{aligned} D_{13}' &= k_{61}P_1P_3 - k_{62}D_{13} && (11)\\ R_5' &= k_{53}X_3 - k_{54}R_5 && (12)\\ P_5' &= k_{15}R_5 - k_{16}P_5 - 2k_{57}P_5^2 + 2k_{58}D_5 && (13)\\ D_5' &= k_{57}P_5^2 - k_{58}D_5, && (14) \end{aligned}$$

and

$$Prob\{X_3 = 0 \text{ in } (t, t+h) | X_3(t) = 0\} = \exp\left(\frac{-D_{13}(t)}{\epsilon}h\right) + o(h),$$

$$Prob\{X_3 = 1 \text{ in } (t, t+h) | X_3(t) = 1\} = \exp\left(\frac{-q}{\epsilon}h\right) + o(h).$$

The parameter q was listed in Table 2. For reasonable values of the other ones, the reader is referred to ([8],[9]).

3.3 The time-averaged deterministic model

As ϵ is (realistically) small, we investigate what happens in the limit $\epsilon \searrow 0$. First, observe that we have uniform (in ϵ) a priori bounds on the derivatives of all the dependent variables; for example, $D_2(t)$ varies uniformly slowly relatively to ϵ, such that it should be legitimate that D_2 be treated as a constant on short time intervals as $\epsilon \searrow 0$.

Renewal reward theory (see [32]) allows us to derive a system of ordinary differential equations replacing (5-10) by a "time-averaged" system in the limit. If D_2 were in fact independent of time, the time average of $X_1(t)$ over

"macroscopic" time intervals (i.e., intervals of scale much larger than ϵ) is $\frac{D_2}{r+D_2}$. The corresponding average of $1 - X_1(t)$ is then $\frac{r}{r+D_2}$. Renewal reward theory implies that this intuition is mathematically accurate.

Specifically, define a cycle to consist of a period of unoccupied time followed by a period of occupied time, ending with detachment. The period of unoccupied time is by construction exponentially distributed with mean ϵ/D_2. Suppose, in the language of renewal reward theory, that no reward is received during this time. The following, occupied part of the cycle is exponentially distributed with mean ϵ/r, and we assume that the reward associated with this period is exactly equal to the amount of occupied time. Renewal reward theory then asserts that the long-term average reward (i.e., the proportion of occupied time) is with probability 1 equal to $E(R)/E(L)$, where $E(R)$ is the expected reward during a cycle and $E(L)$ is the expected length of a cycle.

In the case under consideration

$$E(R) = \epsilon/r, \quad E(L) = \epsilon/r + \epsilon/D_2,$$

so the long-term time average of $X_1(t)$ is $D_2/(r + D_2)$, i.e., $\lim_{\epsilon\to 0} X_{1\epsilon}(t) = \frac{D_2}{r+D_2}$ (here, we denote the random variables X_i as $X_{i\epsilon}$ to emphasize the dependence on ϵ). This time average will hold over any time interval over which D_2 is constant or changes sufficiently slowly. In this time-averaged system Eqns. (5,8) then become

$$R_1' = k_{11}\frac{r}{r + D_2} - k_{12}R_1 \tag{15}$$

$$R_2' = k_{11}\frac{D_1}{s + D_1} - k_{12}R_2 \tag{16}$$

and the remaining equations stay the same. Similarly, Equation (12) becomes

$$R_5' = k_{53}\frac{D_{13}}{(q + D_{13})} - k_{54}R_5.$$

This intuitive argument is not rigorous, but it can be made rigorous. For details, see [8]. We state the result.

Denote by $R_{1\epsilon}, P_{1\epsilon,}D_{1\epsilon}$ etc. the solution of (5-10) for some $\epsilon > 0$ and given initial values $R_1(0), P_1(0), \ldots,$ and denote by R_1, P_1, D_1 etc. the solution of Eqns. (15, 16) ff. for the same initial values.

Proposition 1. *Almost surely for all $t > 0$,*

$$\lim_{\epsilon\to 0} R_{1\epsilon}(t) = R_1(t)$$

$$\lim_{\epsilon\to 0} P_{1\epsilon}(t) = P_1(t)$$
$$\lim_{\epsilon\to 0} D_{1\epsilon}(t) = D_1(t)$$
$$\lim_{\epsilon\to 0} R_{2\epsilon}(t) = R_2(t)$$

etc.

Remarks on the proof. It is easily seen that on any bounded time interval $[0, T]$ the random funcations $\{R_{1\epsilon}, P_{1\epsilon}, \ldots\}$ are an equicontinuous and uniformly bounded set, such that the conditions of the Arzelà-Ascoli Theorem apply. We can therefore extract convergent subsequences, and it suffices then to show that the limits will satisfy the deterministic system (15, 16) ff. This is done by considering the integral version of the systems, partitioning the integrals into Riemann sums with step size Δt, and using the mentioned renewal reward result on each short time interval. For details see [8].

3.4 Interlude: Coupled and Forced Oscillations

The production of the heterodimer D_{13} described in the previous section was the key step leading towards circadian rhythms. These model rhythms are therefore the result of forced oscillations with forcing terms whose frequencies are close but not identical; in other words, we think of circadian rhythms as resonances. This idea is old [34] and somewhat controversial among biologists. In Ref. [9] we presented a detailed discussion arguing in favour of the idea. Rather than repeat this discussion here, we will provide a very brief review of the elementary facts behind resonances, or beats.

Consider the simple coupled system of linear ODEs

$$x_1'' + \omega^2 x_1 = \kappa(x_2 - x_1) \quad (17)$$
$$x_2'' + \omega^2 x_2 = \kappa(x_1 - x_2) \quad (18)$$

where κ is thought of as a small coupling constant, and ω is the natural frequency of a harmonic oscillator. This system is the prototype of coupled oscillators. If we set $Z := x_1 - x_2$, the system can be written equivalently as

$$x_1'' + \omega^2 x_1 = -\kappa Z \quad (19)$$
$$Z'' + (\omega^2 + 2\kappa) Z = 0 \quad (20)$$

so the second equation is now decoupled from the first, and its solution Z can be seen as driving the oscillations of x_1. At least for this linear scenario it is therefore transparent that resonances via coupling and resonances via external forcing are equivalent. Of course, if we set $Y = x_1 + x_2$, the equations

for Y, Z are really decoupled and easily solved. This is a first-year exercise for undergraduate students, with the result $Y(t) = \alpha\cos(\omega t - \beta)$, $Z(t) = A\cos((\sqrt{\omega^2 + 2\kappa})t - \gamma)$. If we abbreviate $a := A - \alpha$, we find

$$x_1(t) = \frac{a}{2}\cos(\omega t - \beta) + \frac{1}{2}A(\cos(\omega t - \beta) + \cos((\sqrt{\omega^2 + 2\kappa})t - \gamma),$$

and after the use of a trigonometric identity

$$\begin{aligned} x_1(t) &= \frac{a}{2}\cos(\omega t - \beta) + A\cos(\frac{1}{2}\left((\omega + \sqrt{\omega^2 + 2\kappa})t - (\beta + \gamma)\right)) \\ &\quad \times \cos\left((\sqrt{\omega^2 + 2\kappa} - \omega)t - \frac{1}{2}(\beta - \gamma)\right) \end{aligned} \tag{21}$$

Recall that κ is assumed to be small relative to ω. Thus

$$\sqrt{\omega^2 + 2\kappa} = \omega + \frac{\kappa}{\omega} + O(\kappa^2),$$

and the last term in (21) produces an envelope $\cos\left(\frac{\kappa}{\omega}t - \ldots\right)$ with the long "circadian" period $T_2 = \frac{2\pi\omega}{\kappa}$, as compared to the shorter "ultradian" period $T_1 = \frac{2\pi}{\omega}$. Note that the same T_2 is produced by constant quotients of ω and κ; hence circadian rhythms may be produced by various choices of ultradian oscillators and coupling strengths.

The (stochastic or time-averaged) systems considered earlier are more complex than the simple harmonic oscillators just described, but the basic mechanism is very similar. In fact, linear bifurcation analysis can be applied to understand the origins of the oscillations at the TTO level, and beyond this the secondary oscillations reduce to sinusoidal waves via Fourier series expansions. We discuss this in the next section.

3.5 Ultradian oscillations from a Hopf bifurcation

Two questions arise naturally from what has been said so far. They are

- What are the principal reasons causing the primary oscillations?
- Is it really necessary to introduce secondary oscillations from couplings between primary oscillators to obtain circadian frequencies? Is it not rather possible to obtain circadian rhythms directly from TTOs, with reasonable ranges of parameter values based on known biochemistry?

The second question was investigated to some extent in Ref. [8]; while there are parameter ranges that can stretch the periods of TTOs to circadian range, these look biologically not realistic.

In this section we address the first question and show that the primary oscillations arise from a Hopf bifurcation. This mechanism also provides the methodology (at least in principle) to predict periods of the primary oscillations, at least while the parameters are near the bifurcation points.

Consider the time-averaged single primary oscillator

$$\begin{aligned}
\frac{dR_1}{dt} &= k_{11}\frac{r}{r+D_2} - k_{12}R_1 \\
\frac{dP_1}{dt} &= k_{13}R_1 - k_{14}P_1 - 2k_{15}P_1^2 + 2k_{16}D_1 \\
\frac{dD_1}{dt} &= k_{15}P_1^2 - k_{16}D_1 \\
\frac{dR_2}{dt} &= k_{11}\frac{D_1}{s+D_1} - k_{12}R_2 \\
\frac{dP_2}{dt} &= k_{23}R_2 - k_{14}P_2 - 2k_{25}P_2^2 + 2k_{26}D_2 \\
\frac{dD_2}{dt} &= k_{25}P_2^2 - k_{26}D_2
\end{aligned} \tag{22}$$

To simplify a systematic investigation of the dependence of the periods on the parameters, let $k_{23} = k_{13}, k_{25} = k_{15}, k_{26} = k_{16}$, and linearize the system about its unique positive equilibrium $(R_{1E}, P_{1E}, D_{1E}, R_{2E}, P_{2E}, D_{2E})$ (it is easy to see that there is such an equilibrium). The linearization yields the 6-by-6 matrix

$$A = \begin{pmatrix}
-k_{13} & 0 & 0 & 0 & 0 & \frac{-k_{11}r}{(r+D_{2E})^2} \\
k_{13} & -k_{14} - 2k_{15}P_{1E} & 2k_{16} & 0 & 0 & 0 \\
0 & 2k_{15}P_{1E} & -k_{16} & 0 & 0 & 0 \\
0 & 0 & \frac{k_{11}s}{(s+D_{1E})^2} & -k_{12} & 0 & 0 \\
0 & 0 & 0 & k_{13} & -k_{14} - 2k_{15}P_{2E} & 2k_{16} \\
0 & 0 & 0 & 0 & 2k_{15}P_{2E} & -k_{16}
\end{pmatrix}$$

Its eigenvalues satisfy $\det(A - \lambda I) = 0$, which yields the characteristic equation

$$(k_{12}+\lambda)^2(k_{16}+\lambda)^2(k_{14}+2k_{15}P_{1E}+\lambda)(k_{14}+2k_{15}P_{2E}+\lambda) + c_0^2 = 0. \tag{23}$$

Here,

$$c_0^2 = \frac{4rsk_{13}^2k_{15}^2k_{11}^2}{(s+D_{1E})^2(r+D_{2E})^2}.$$

Note that the numerator is a simple product of parameters, but the dependence of c_0 on the parameters is somewhat more complex because the equilibria values D_{1E}, D_{2E} also depend on the parameters. To identify oscillatory solutions with (long) periods we need a complex conjugate pair of eigenvalues with positive real part and (small) imaginary parts. Clearly, $c_0^2 = 0$ in (23) produces 6 real and negative eigenvalues (counted with their multiplicity). If we now increase c_0^2, one pair of eigenvalues approaches and eventually crosses the imaginary axis, producing the Hopf bifurcation and the oscillations.

The only way to force the crossing of the imaginary axis at small imaginary value is to move a pair of eigenvalues closer to the imaginary axis to begin with (i.e., when $c_0^2 = 0$). Numerical experiments to this end are given in [8].

3.6 On numerical and real experiments

In Ref. [8] we presented several simulations performed with the XPPAUT package (see [11], or http://www.math.pitt.edu/ ~bard/xpp/xpp.html). The parameters used in these studies are those from Table 1. Here we present two graphs associated with the deterministic (time-averaged) model.

First, Figure 4 shows the time evolution of the numbers of the proteins P_1 and P_3. They oscillate with a period of about 3 hours but differ slightly in their periods (of course, this was set up to be so). A slight circadian variation is seen, but is much more promiment in Figure 5, where the responses of the protein products of the fifth DNA site are shown; note the time lag of D_5 with respect to D_{13}.

In Ref. [8] the same calculation was done for the stochastic model. We used Gillespie's method [12], where the ϵ was chosen as $2.8 \times 10^{-5} hrs$. The results are essentially identical to the ones for the time-averaged model.

As a control measure we performed some calculations with larger ϵ, for example $\epsilon = 2.8 \times 10^{-3}$ hrs and $\epsilon = 0.028$ hrs. For the former case, especially, the results were still very close to the time-averaged simulations. For the latter case deviations from the time-averaged simulations became noticable: in particular, the amplitude of the circadian oscillations in D_5 fluctuated stochastically and their period decreased slightly.

The circadian period is remarkably robust with respect to the choice of ϵ. We demonstrate this in [8] by computing Fourier power spectra of the D_5

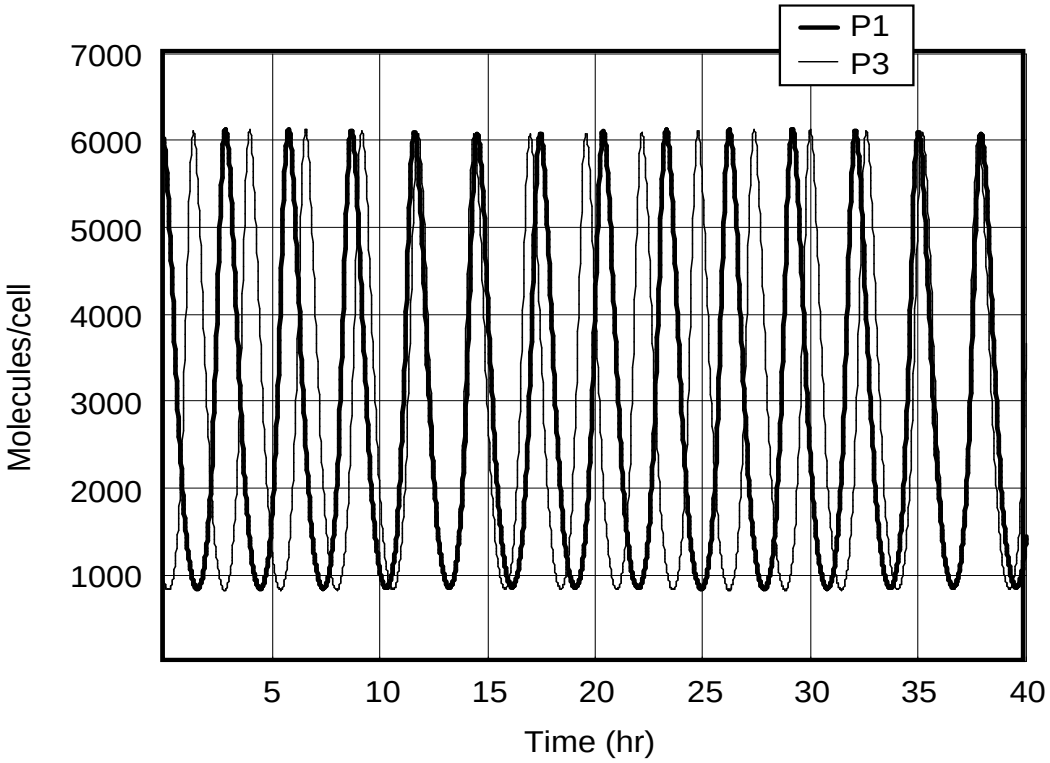

Figure 4: The time evolution of the proteins P_1 and P_3 according to the time-averaged model

time series generated by simulations with $\epsilon = 2.8 \times 10^{-5}$ and $\epsilon = 2.8 \times 10^{-2}$. The reader is referred to [8] for details.

Fourier power spectra offer a natural way of investigating whether our model applies to reality. Specifically, efforts are currently under way (with suitable plants) to observe periodic activity levels in the absence of external circadian stimuli. Power spectra of the collected time series display (of course) a strong circadian peak, but there is evidence of a (much weaker) power concentration at ultradian frequency levels. This is consistent with our theory, but the data are of insufficient resolution to decide whether couplings between ultradian oscillators are responsible for the circadian rhythm. More experimental work on this is in progress.

3.7 On Kolmogorov master equations: an example

The theory presented in the above sections is incomplete in the sense that there is no mechanism to analytically investigate fluctuations caused by the

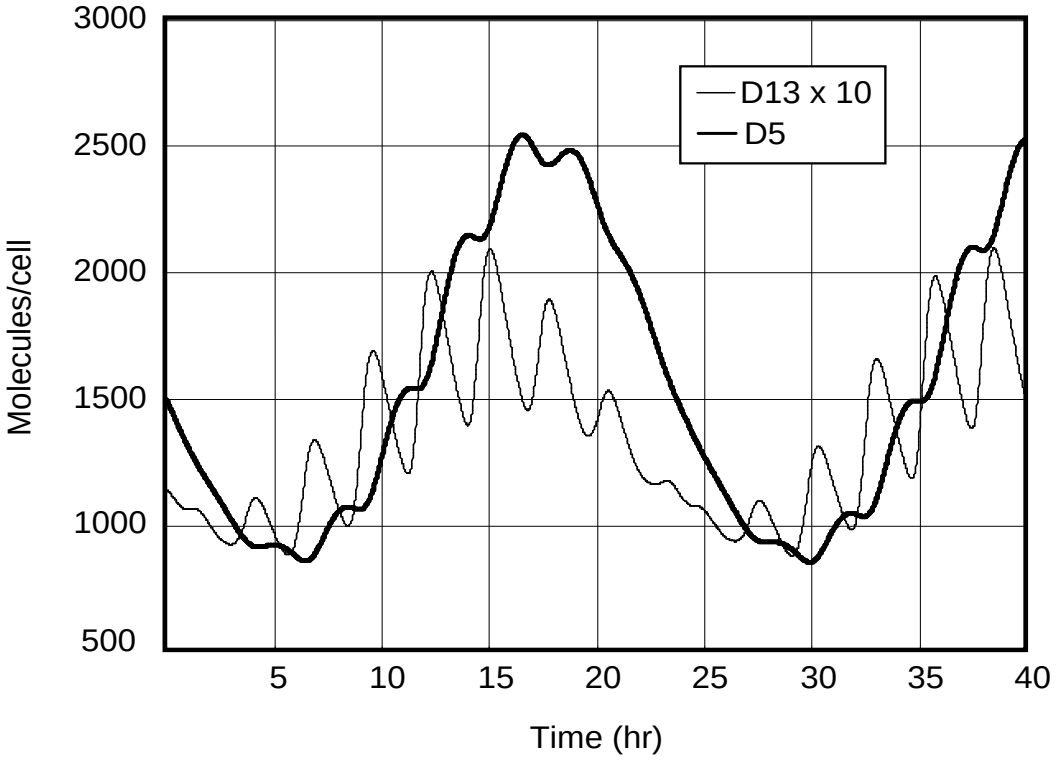

Figure 5: The time evolution of the heterodimer D_{13} and the homodimer D_5 according to the time-averaged model

random variables X_1, Y_1 etc., and the associated parameter ϵ, or by stochastic fluctuations of the initial state of the system (which, for the problem under consideration, should be completely irrelevant). All we proved (in Proposition 1) is that for given fixed initial values the fluctuations will indeed vanish as $\epsilon \to 0$.

The method of choice for gaining additional information is to set up a system of partial differential equations (the Kolmogorov master equations, which we will for short call KME) for the probability densities that the random variables $X_1, \ldots$ at time t and the dependent variables $R_i(t), P_i(t), D_i(t)$ assume certain values. This is possible but somewhat cumbersome for the full system starting with equations (5)-(10). Note that there are four possible right hand sides for the equations, each associated to whether or not X_1 and Y_1 are 0 or 1, and each assumed with certain probabilities. Instead of deriving the KME for the full complicated system we shall discuss the following much simpler example.

Let $X = X(t)$ be a random variable alternating between 0 and 1, and let $N = N(t)$ be a dependent variable whose evolution is driven by the law

$$\dot{N} = \begin{cases} f_1(N) & \text{if } X = 1 \\ f_0(N) & \text{if } X = 0 \end{cases}$$

f_0 and f_1 could really be any (smooth) functions but later we will set $f_1(N) = 1 - N$, $f_0(N) = -N$, or more compactly $f_i(N) = i - N,\ i = 0, 1$. This choice implies that the values of N will remain confined to $[0, 1]$ if they are so at time 0.

Now assume that there are probability densities $p_i(N, t)$ such that for any Borel set $I \subset [0, 1]$ and $i = 0, 1$

$$Prob\{N(t) \in I, X(t) = i\} = \int_I p_i(N, t)\ dN.$$

Our objective is to find (and use) evolution equations (the KME) for the p_i. We need two fundamental tools to this end:

1. The definition of conditional probabilities:

$$Prob\{A \cap B\} = Prob\{A|B\}Prob\{B\}$$

2. The definition (and existence) of transition rates for X.

For the subsequent calculation we denote by $T_i(t)N_0$ the unique solution of $\dot{N} = f_i(N),\ N(0) = N_0$. Also, we consider a time interval $[t, t + \Delta t]$, and consider

$$\begin{aligned} &Prob\{N(t + \Delta t) \in I, X(t + \Delta t) = 1\} = \\ &\quad Prob\{N(t + \Delta t) \in I, X(t + \Delta t) = 1, X(t) = 1\} \\ &+ \quad Prob\{N(t + \Delta t) \in I, X(t + \Delta t) = 1, X(t) = 0\} \\ &= \quad Prob\{N(t) \in T_1(-\Delta t)I, X(t + \Delta t) = 1, X(t) = 1\} \\ &+ \quad Prob\{N(t) \in T_0(-\Delta t)I, X(t + \Delta t) = 1, X(t) = 0\} + o(\Delta t). \end{aligned}$$

In the last identity it has tacitly been assumed that Δt is small and that two flips of X in $[t, t + \Delta t]$ are of probability $o(\Delta t)$; this is the meaning of "$\approx$". Also, the last line assumes that the flip $0 \to 1$ of X happens at exactly time $t + \Delta t$; this assumption is too simplistic, but (as the reader may verify) the subsequent calculations will produce the same final result of the flip is assumed to occur at $t + s\Delta t$ for some $s \in (0, 1]$. We continue the calculation by converting to conditional probabilities:

$$
\begin{aligned}
&Prob\{N(t+\Delta t) \in I, X(t+\Delta t) = 1\} = \\
&\qquad Prob\{X(t+\Delta t) = 1 | X(t) = 1, N(t) \in T_1(-\Delta t)I\} \\
&\qquad \times Prob\{X(t) = 1, N(t) \in T_1(-\Delta t)I\} \\
&+ \quad Prob\{X(t+\Delta t) = 1 | X(t) = 0, N(t) \in T_0(-\Delta t)I\} \\
&\qquad \times Prob\{X(t) = 0, N(t) \in T_0(-\Delta t)I\} + o(\Delta t).
\end{aligned}
$$

We next invoke the second key ingredient, namely the assumption that there are *transition rates* $\lambda > 0$ and $\mu > 0$ such that

$$
\begin{aligned}
Prob\{X(t+\Delta t) = 0 \quad | X(t) = 1, N(t) \in T_1(-\Delta t)I\} &= \mu \Delta t + o(\Delta t) \\
Prob\{X(t+\Delta t) = 1 \quad | X(t) = 0, N(t) \in T_0(-\Delta t)I\} &= \lambda \Delta t + o(\Delta t).
\end{aligned}
$$

We first assume that λ and μ are constant: we call this the *constant* case; the important case where λ and μ depend on the sets $T_0(-\Delta t)I, T_1(-\Delta t)I$, respectively, and assume limits $\lambda(N), \mu(N)$ as these sets concentrate on $N(t)$ is called the *variable* case. The variable case is the important one for the TTO examples; we will explore the example $\lambda(N) = N, \mu = const.$ in more detail.

For the case where λ and μ are constant, the previous calculations readily lead to

$$
\begin{aligned}
\int_I p_1(N, t+\Delta t)\ dN = \quad & \textstyle\int_{T_1(-\Delta t)I} p_1(N,t)\ dN \cdot (1 - \mu\Delta t + o(\Delta t)) \\
& + \textstyle\int_{T_0(-\Delta t)I} p_0(N,t)\ dN \cdot (\lambda \Delta t + o(\Delta t)) \qquad (24)
\end{aligned}
$$

By using the simple substitution $N = z - \Delta t f_1(z)$ we proceed to compute $dN = (1 - \Delta t f_1'(z))dz$ and

$$
\int_{T_1(-\Delta t)I} p_1(N,t)\ dN = \int_I p_1(z - \Delta t f_1(z), t)(1 - \Delta t f_1'(z))\ dz + o(\Delta t).
$$

Substituting this into (24), collecting terms and concentrating I to the point N leads to

$$
\begin{aligned}
&p_1(N, t+\Delta t) - p_1(N,t) = \\
&\qquad p_1(N - f_1(N)\Delta t, t)(1 - f_1'(N)\Delta t)(1 - \mu \Delta t) \\
&\qquad -p_1(N,t) + p_0(N - f_0(N)\Delta t, t)(1 - f_0'(N)\Delta t)\lambda \Delta t + o(\Delta t),
\end{aligned}
$$

and after division by Δt and taking the limit as $\Delta t \to 0$, we find

$$\begin{aligned} \partial_t p_1 + \partial_N (p_1 f_1) &= \lambda p_0 - \mu p_1 \\ \partial_t p_0 + \partial_N (p_0 f_0) &= -\lambda p_0 + \mu p_1. \end{aligned} \tag{25}$$

This system of equations is known as the Kolmogorov master equations (short, KME) for the original equation.

Remarks.

- One easily sees that the same system (25) emerges if the "flip" of X from 0 to 1 occurs at some time $t + s\Delta t,\ s < 1$ (rather than at $t + \Delta t$, as assumed above).
- If λ and μ depend on the previous state of the system, the equations (25) emerge with $\lambda = \lambda(N)$ and $\mu = \mu(N)$ on the right-hand side.
- Of particular interest for us is what happens in the "fluid-dynamic" limit where $\lambda \to \frac{\lambda}{\epsilon}, \mu \to \frac{\mu}{\epsilon}$ and $\epsilon \to 0$. We address this, as is common in theories of kinetic equations, by studying the associated moment equations.

3.8 Moment equations

Consider the special case of (25) where $f_i(N) = i - N$. The rescaled equations become

$$\begin{aligned} \partial_t p_1 + \partial_N ((1-N) p_1) &= \frac{1}{\epsilon}(\lambda p_0 - \mu p_1) \\ \partial_t p_0 + \partial_N (N p_0) &= \frac{1}{\epsilon}(\mu p_1 - \lambda p_0). \end{aligned} \tag{26}$$

and we complement these with the boundary conditions $p_1(0,t) = p_0(1,t) = 0$.

We study the moment equations associated with (26) for the constant case and for the case where μ is constant but $\lambda = \lambda(N) = N$. Denote for $i = 0, 1$

$$\begin{aligned} P^i(t) &= \int_0^1 p_i(N,t)\ dN, \quad P(t) = P^0(t) + P^1(t) = 1 \\ \mathbb{E}^i(t) &= \int_0^1 N p_i(N,t)\ dN, \quad \mathbb{E}(t) = \mathbb{E}^1(t) + \mathbb{E}^2(t) \\ \mathbb{E}_2^i(t) &= \int_0^1 N^2 p_i(N,t)\ dN, \quad \mathbb{E}_2(t) = \mathbb{E}_2^0(t) + \mathbb{E}_2^1(t), \text{etc., and} \\ V(t) &:= \mathbb{E}_2(t) - \mathbb{E}^2(t) \text{ (the variance).} \end{aligned}$$

By multiplying (26) with suitable powers of N, integration and by using the boundary conditions, we obtain the moment equations. In the case where both λ and μ are constant, these are

$$\dot{P^1} = \frac{1}{\epsilon}(\lambda P^0 - \mu P^1)$$
$$\dot{P^0} = \frac{1}{\epsilon}(\mu P^1 - \lambda P^0)$$

with the solution

$$P^1(t) = \frac{\lambda}{\lambda+\mu} + \left(P^1(0) - \frac{\lambda}{\lambda+\mu}\right) e^{-\frac{\lambda+\mu}{\epsilon}t}, \quad P^0(t) = 1 - P^1(t). \qquad (27)$$

Clearly, $P^1(\infty) = \frac{\lambda}{\lambda+\mu}$, $P^0(\infty) = \frac{\mu}{\lambda+\mu}$, and the solution converges $\frac{1}{\epsilon}-$ exponentially fast towards these steady values. For the next moment, we get the system

$$\begin{aligned} \dot{\mathbb{E}^1} &= P^1 - \mathbb{E}^1 + \frac{1}{\epsilon}(\lambda \mathbb{E}^0 - \mu \mathbb{E}^1) \\ \dot{\mathbb{E}^0} &= -\mathbb{E}^0 + \frac{1}{\epsilon}(\mu \mathbb{E}^1 - \lambda \mathbb{E}^0) \end{aligned}$$

and $\dot{\mathbb{E}} = P^1 - \mathbb{E}$; after times of order O(1) the latter becomes

$$\dot{\mathbb{E}} = \frac{\lambda}{\lambda+\mu} - \mathbb{E},$$

which should be compared with the original equation $\dot{N} = X - N$: assuming constant switching rates and the scaling under consideration, the renewal reward result used in Proposition 1 maybe applied to the current setting and leads to the same equation for $\mathbb{E}$.

Note that the equation for $\mathbb{E}$ contains ϵ only via P^1, and we saw that this $\epsilon-$ dependence vanishes after times of order O(1) (actually, even faster). The equations for $\mathbb{E}^i$ have no simple limits as $\epsilon \to 0$, but it is not hard to compute the $\mathbb{E}^i$'s explicitly. For example, one finds that up to errors of order ϵ in the first two terms

$$\mathbb{E}^1(t) = \left(\frac{\lambda}{\lambda+\mu}\right)^2 + \frac{\lambda}{\lambda+\mu}\left(\mathbb{E}(0) - \frac{\lambda}{\lambda+\mu}\right) e^{-t} + Ce^{-(1+\frac{\lambda+\mu}{\epsilon})t},$$

and it is clear from this representation how $\mathbb{E}^1$ decomposes into a steady part and parts decaying on different time scales.

As for the variance V, similar calculations show that

$$\dot{V} = 2(\mathbb{E}^1 - P^1\mathbb{E}) - 2V; \tag{28}$$

substituting the obtained formulas for $\mathbb{E}^1, P^1$ and $\mathbb{E}$ shows readily that the first term on the right of (28) is of order $O(\epsilon)$ for times $O(1)$ and vanishes completely for such times if $\epsilon = 0$. This provides a qualitative means to control the variance of the distribution of N after an initial layer that may be caused by initial uncertainty in N.

The variable case is much harder, but we first note that all required information is implicitly contained in the linear system of PDEs (26). If we could solve this system explicitely the solution would provide anything we wish to know. We could, of course, rely on numerical solutions, but keep in mind that we are interested in what happens in the limit $\epsilon \to 0$, a most delicate limit from a numerical point of view. The system of moment equations, which provided such a drastic simplification in the *constant* case, is now far less friendly because it is not a closed system: for the example we obtain

$$\begin{aligned}
\dot{P}^1 &= \frac{1}{\epsilon}(\mathbb{E}^0 - \mu P^1) \\
\dot{P}^0 &= \frac{1}{\epsilon}(\mu P^1 - \mathbb{E}^0)
\end{aligned}$$

so $\mathbb{E}^0$ is required to solve the system for P^1, and then $P^0 = 1 - P^1$. For the $\mathbb{E}^i$ we have

$$\begin{aligned}
\dot{\mathbb{E}}^1 &= P^1 - \mathbb{E}^1 + \frac{1}{\epsilon}(\mathbb{E}^0_2 - \mu\mathbb{E}^1) \\
\dot{\mathbb{E}}^0 &= -\mathbb{E}^0 + \frac{1}{\epsilon}(\mu\mathbb{E}^1 - \mathbb{E}^0_2)
\end{aligned}$$

and, as before, $\dot{\mathbb{E}} = P^1 - \mathbb{E}$.

The equation for the variance $V := \mathbb{E}_2 - \mathbb{E}^2$ remains identical to the "constant" case, as in (28):

$$\frac{d}{dt}V = 2(\mathbb{E}^1 - P^1\mathbb{E}) - 2V.$$

A rapid count shows that at this level we have 4 equations for the 5 unknowns

$$P^1, \mathbb{E}^0, \mathbb{E}^1, \mathbb{E}^0_2 \text{ and } V$$

(we can ignore (the equation for) P^0 because $P^0 + P^1 = 1$). We can now either include additional higher order moments (leading to a larger and still

not closed set of equations) or use a closure relation to eliminate one of the variables. This is, of course, exactly the question of closing hydrodynamic equations emerging as moment equations from a kinetic equation.

To complete this section we will show a simple (and non-rigorous) way of obtaining closed limits of moment equations based on the following two steps.

- Multiply the equation for P^1 by ϵ, then set $\epsilon = 0$. This suggests that $\mathbb{E}^0 \approx \mu P^1$. This is a fluid dynamic approximation in the sense of kinetic theory.
- Assume that $\mathbb{E}^0 \approx P^0 \mathbb{E}$. This is an (asymptotic) independence assumption.

These two steps together, if valid, imply that

$$P^1 = 1 - P^0 = 1 - \frac{\mu}{\mathbb{E}} P^1$$

such that $P^1 = \frac{\mathbb{E}}{\mathbb{E}+\mu}$, and therefore

$$\dot{\mathbb{E}} = \frac{\mathbb{E}}{\mathbb{E}+\mu} - \mathbb{E}.$$

The reader should compare this to the constant case, and to our time-averaged systems of TTO equations. Notice also that under the two assumptions, we necessarily find $\mathbb{E}^1 = P^1 \mathbb{E}$, and the variance equation therefore becomes

$$\dot{V} = -2V,$$

meaning that whatever variance is present must have come from initial uncertainty and will diminish exponentially fast with time.

At this level we have not used the moments $\mathbb{E}_2^i$ at all.

Of course, our two closure assumptions remain to be justified; one way to pursue this is by using the renewal reward argument which we employed to obtain the time-averaged version of the TTO equations, adapted to the present example. The details of this are still being worked out as I write this.

For more sophisticated (higher order) closures we should emulate Grad's closures or related approaches (see [31]). I am not aware of any applications of this methodology to the current biological context, but the importance of good control of the size of stochastic fluctuations is clear.

4 Fokker-Planck Type Models for Multilane Traffic Flow

The various subjects covered in this article (and in my Porto Ercole lectures) differ widely in terms of their origins — physical chemistry, microbiology, traffic flow on roads — but they share averaging concepts as a common thread. In the crystallization problems discussed in the first section, we averaged over many small Voronoi cells; in the modeling of TTOs time-averaging over times of order larger than ϵ led to simpler systems of evolution equations for the molecular concentrations in a cell; in this present section we shall use the idea of (heuristic) ensemble averaging to set up kinetic equations for multi-lane traffic flow. We begin with a short review of what types of traffic models are being studied.

4.1 Types of traffic models, and fundamental diagrams

Most of the traffic models in use for practical applications belong to one of the following three categories:

1. Microscopic ("follow-the leader") models, in which each individual car is modeled by its own differential-delay equation, of type

$$\ddot{x}_i = C(v_i)\frac{(v_{i+1} - v_i)^{m_1}(t-\tau)}{(x_{i+1} - x_i)^{m_2}(t-\tau)},$$

where x_i and v_i denote the position and speed of the i-th car (counting in the direction of traffic flow), x_{i+1} and v_{i+1} correspond to the leading car, and τ is the individual reaction time. The factor $C(v_i)$ is usually (but not always) taken as a constant, and the powers m_1, m_2 vary from model to model; in the simplest case $m_1 = m_2 = 1$. The choice $m_1 = 2, m_2 = 1$ makes $C(v_i)$ dimensionless. As presented, these models assume single-lane traffic; even two-lane traffic introduces serious complications, because lane-changing has to be taken seriously and requires careful bookkeeping on car-interactions between the lanes.

Microscopic models can be and have been used to compute density-flux relationships ("fundamental diagrams") in equilibrium situations, for example in carefully monitored flow through a tunnel where all vehicles keep approximately the same distance and speed; one may define a density $\rho(x_i) = \frac{1}{x_i - x_{i+1}}$ and an average speed $u_i(x_i) = v_i$ and compute a fundamental diagram $u = u(\rho)$ from basic observational facts such as $u(\rho) = v_{max}$ (the speed limit) for $\rho \in [0, \rho_{crit})$ (a critical density,

identified by observations, at which drivers will begin to slow down), $u(\rho_{max}) = 0$, and integration of the model equations.

2. At the other end of the spectrum are completely macroscopic models, which are PDEs, typically (systems of) conservation laws relating density and flux. For example, for $\rho = \rho(x,t)$ and $j = j(x,t) = \rho(x,t)u^e(x,t)$

$$\partial_t \rho + \partial j = 0.$$

Here, u^e and ρ are linked by a (presupposed) fundamental diagram, such that the equation becomes closed. Models of this type were already considered by Lighthill and Whitham 40 years ago; as scalar conservation laws they enjoy attention as good examples for the formation and propagation of shock and rarefaction waves. They also are in use for real highway simulations and give satisfactory results for many (but not all) traffic situations. The underlying assumption is that the traffic is close to equilibrium all the time, and that the equilibria are stable. Unfortunately, this does not always seem to be true. We discuss these weaknesses and suggested remedies below.

3. The third class of traffic models, whose interpretation requires a statistical point of view, are *kinetic* models. Such models provide partial differential equations of transport (or drift-diffusion) type with interaction terms for car density functions $f(x,v,t)$ so that macroscopic density ρ and flux j are given as moments $\rho(x,t) = \int f(x,v,t)\, dv$ and $j(x,t) = \int vf(x,v,t)\, dv$. We will mention examples of such models in the sequel and discuss Fokker-Planck type models for multilane flow in some detail.

4.1.1 Fundamental diagrams

As mentioned, a fundamental diagram is a relationship between ρ and j (or ρ and $u = j/\rho$) in equilibrated traffic. It is a priori not even clear that such a relationship should exist; theoretically, any different average speed up to the speed limit (and possibly beyond) is conceivable: in other words, the measures $\rho\delta(v-u)$, for which all drivers simply move at the same (possibly large) speed u should from the outset not be ruled out as solutions. In fact, such "synchronized traffic" is observed and even desirable on freeways, and large ρ and u together produce desired large flux. The problem is that such values for ρ and u are also inherently very dangerous, and so drivers will reduce ρ by increasing their distance from the lead car.

Traffic observations seem to produce acceptable functional relationships consistent with the existence of a fundamental diagram in many situations. However, there are a few exceptions, which seem to require refined theories.

A first regime where the fundamental diagram is easily obtained is when the road (or the lanes in case of two- or multi-lane traffic) is almost empty. It is reasonable to assume that (almost) everyone will drive at the speed limit v_{max}, such that $u = v_{max}$ and u is a constant function of ρ for small ρ.

For really high densities (say, two car lengths between any two cars) drivers whould be forced to reduce their speed; it is, however, not clear what the end speed should be — that may well be dependent on the society under consideration. It is not even clear whether there would be a stable steady state: moving jams are observed in situations like this, indicating that a steady flux associated with such a high density might be unstable. We refer to [23] for observations and theories on this regime. For the maximal density (bumper-to-bumper traffic) one expects standing traffic, and hence zero flux.

The third regime is the intermediate domain, and this is where things depend on the type of road. On a single-lane road the fundmental diagram seems to make good sense and produce a decreasing average speed as a function of density; on multi-lane highways, however, measurements seem to suggest that there is a density domain $\rho \in [\rho_1, \rho_2]$ where the fundamental diagram is multivalued; this is reported in the papers [23], [24] and elsewhere. As the phenomenon occurs only on multi-lane highways, it is clear that it must be related to lane-changing.

4.1.2 Some examples

Given the uncertainties surrounding the existence and stability of fundamental diagrams, the need for more sophisticated models was there a long time ago. A natural generalization was to replace scalar conservation laws by systems, where a separate equation for u is assumed. An early model of this type was due to Payne and Whitham, and reads

$$\begin{aligned} \rho_t + (\rho u)_x &= 0 \\ u_t + uu_x + \frac{1}{\rho} a_{PW}(\rho)\rho_x &= \frac{1}{T^e(\rho)}[u^e(\rho) - u]. \end{aligned} \tag{29}$$

The idea here is that the average speed is assumed to respond to density gradients via a scaled "anticipation" function a_{PW}, and that traffic always tries to relax to an equilibrium speed $u^e(\rho)$ (thus effectively postulating a fundamental diagram) on a characteristic time scale $T^e(\rho)$.

This model is unsatisfactory from several points of view. First, it contains three a priori unknown functions whose existence is debatable. Second, as shown by Daganzo [4], it may produce the nonsensical phenomenon that traffic gradients may force $u(x,t) < 0$, i.e., traffic may flow backwards. Indeed, something not seen in reality. A moment's thought about the structure of equation (29) will convince the reader that strong positive gradients of ρ may force rapid and unchecked decrease in u, so much so that u may become negative.

A. Aw. and M. Rascle suggested a modification of the model which avoids the pathology:

$$\begin{aligned} \rho_t + (\rho u)_x &= 0 \\ u_t + u u_x + \rho \partial_\rho (u_{AR}(\rho)) u_x &= \frac{1}{T^e(\rho)}[u^e(\rho) - u]. \end{aligned} \tag{30}$$

This system suggests a transport equation for u, so that positivity of u is ensured. However, a fundamental diagram is still implicitly assumed. We note in passing that Klar and Rascle [1] provided a derivation of this equation from microscopic follow-the-leader models. If $T^e = \infty$ and we abbreviate $p = u_{AR}$ then the second equation can be rewritten as

$$(u + p(\rho))_t + u(u + p(\rho))_x = 0.$$

As is obvious from this version, p has the dimension of a speed. Recent analytical work on solving Riemann problems for this model employs this version. (see [17]).

4.2 On kinetic models

Kinetic models for traffic flow are almost as old as theoretical traffic studies themselves. We refer to [20] for a listing. We begin our brief discussion of kinetic models with a list of desirable properties of "good" kinetic models; some of these properties are easily understood and implemented into models. Other properties are desirable, but very elusive.

1. The model should incorporate realistic scales for quantities like speed, acceleration, density, flux, and so on. While this sounds eminently reasonable, it rules out certain types of models from the outset, as will be shown below.

2. The fundamental diagram should (in principle) be computable; on multi-lane highways, where lane-changing is an option, it should display multi-valued regimes consistent with observations.

3. If diffusive mechanisms are absent in or removed from the model (assuming that is possible), the model should possess "trivial" equilibria $\rho\delta_u(v)$, i.e., equilibria which correspond to states where all drivers keep constant distance to the lead car and drive at the same speed u. There are traffic scenarios where such equilibria are observed, for example, in moderately dense traffic, with no lane-changing, on freeways. These scenarios are referred to as "synchronized" traffic in [23].

4. The model should predict traffic phenomena like stop-and-go waves or traffic synchronization.

5. Ideally, the model should be amenable to "validation" from microscopic dynamics via the formulation of a Liouville equation for an "N-" car probability density on a (multi-lane) highway, the establishment of a hierarchy of equations for the reduced correlation functions arising as marginals, and a closure relation analogous to the molecular chaos hypothesis in particle dynamics. However, this last point is likely the most elusive of them all, because there is no reason why there should be such a thing as "vehicular chaos", although attempts to overcome this obstacle have been made (see [25]). This is in stark contrast with equations for rarefied gases or plasmas like the Boltzmann or Vlasov equations, where such validations have been performed.

We will discuss two examples of kinetic models: a model of Boltzmann-Enskog type due to Klar and Wegener ([25], [26]) in which much effort was devoted to point 5 above; and second, a class of models of Fokker-Planck type, due to Illner, Klar and Materne ([20]), for which the 5th item above remains rather elusive, but the other four are attainable.

4.2.1 Enskog-type models: description and critique

Our discussion of these models will be cursory and incomplete; it will be argued that their value is mostly of instructional nature, because, as we state here from the outset, they belong to a (large) class of kinetic models which does not satisfy point 1 from above.

Following the notation from [25] we consider a freeway of N lanes and denote by $f_\alpha = f_\alpha(x, v, t)$ the kinetic traffic density on the $\alpha-$th lane, $\alpha = 1, \ldots, N$. Further $x \in \Re, v \in [0, v_{max}], t \in [0, \infty]$. The equation for f_α reads

$$\partial_t f_\alpha + v\partial_x f_\alpha = C_\alpha(f_1^{(2)}, \ldots, f_N^{(2)}, f_1, \ldots, f_N) \tag{31}$$

where $C_\alpha(\ldots)$ denotes the interaction terms (sometimes also called "collision terms", but this terminology is unfortunate for the application at hand) leading to gains or losses for f_α in the space-speed-time domain $[x, x+dx] \times [v, v+dv] \times [t, t+dt]$. As indicated, the interaction terms may depend (in complicated ways) on the one- and two-vehicle density distributions on various lanes; for example, the gain and loss terms due to braking contained in C_α will be terms

$$(G_B - L_B)(f_{\alpha-1}, f_\alpha^{(2)}, f_{\alpha+1}),$$

where it is indicated that the dependencies will typically involve two-car densities on the same lane (completely reasonable, because you will react relative to what the lead car does), and the one-car densities on adjacent lanes (because these will impact your lane-changing behaviour). In addition to terms due to braking, there will be also terms due to acceleration, gains and losses due to lane changes to and from the lane on the left, and gains and losses due to lane changes to and from the lane to the right. We will not bother the reader with detailed presentations of all these terms; only one, the G_B, is reproduced here for instructional purposes:

$$\begin{aligned} G_B(x,v,t) \;=\; & \int_{v_+} \int_{w>v_+} P_B(w, v_+, f_{\alpha+1}(x, \ldots)) \\ & \sigma_B(w \to v; v_+)|w - v_+|\; f_\alpha^{(2)}(x, w, x + H_0 + T_B w, v_+)\; dw\; dv_+. \end{aligned}$$

For the sake of brevity we refrain from discussing the complete terminology used here; suffice it to say that $P_B(\ldots)$ denotes a braking probability, which will depend on the speed w of the considered car at x relative to the speed of v_+ of the leading car, at $x + H_0 + T_B w$. $\sigma_B(\ldots)$ is a probability to brake to v given that the previous speed was w, and given that the lead car moves at v_+. $\sigma_B(\ldots)|w - v_+|$ then assumes the part of a "collision" (or better, interaction) kernel.

We refer the reader to [25] for the structure of all the other terms arising in the Enskog collision operator. Our main purpose here is to see how these models stand up to the above wish list of properties 1-5.

In passing, I must state here that these Enskog models were my personal introduction to research on kinetic traffic models; specifically, the question of how to use them to calculate fundamental diagrams presented challenging analytical problems inasmuch as the existence, uniqueness and structure

of equilibria (from which the $(\rho, j)-$ values on the fundamental diagram are calculated) is highly nontrivial for these equations. Equilibria were first computed by numerical means alone; later, in Refs. ([21],[22]) analytical existence and uniqueness results were proved for special cases.

In these efforts it became transparent that the equilibria will significantly vary with the choice of the probability densities σ_B. And this is weakness no. 1 of the models: it is a priori not clear how σ_B should be chosen! As explained in the references, the logical choice $\sigma_B(w \to v; v_+) = \delta(v - v_+)$ leads to total cancellation of the interaction terms; hence Klar and Wegener used $\sigma_B(\ldots) = \frac{1}{w(1-\beta)}\chi_{[\beta w, w]}(v)$, where $\beta \in (0,1)$ is a parameter. But other choices are possible and will change the equilibria, and this puts a question mark to the computability of the fundamental diagram, point 2 in our list.

Second, we point out that (31) needs a closure relation before being of use; as written, one needs to know two-vehicle correlation functions to compute one-vehicle densities. The idea of closure relations is to express the former in terms of the latter, by a more or less heuristic factorization including correlation factors (as the standard molecular chaos assumption is certainly not true). We omit the details, but the closure relation problem is closely associated with the 5th item on our list.

The third, final and most serious objection is related to the first one. Why, in fact, is it that we have the freedom of choice for σ_B ? If we had elastic particles on the real line, moment and energy conservation would dictate that the only possibility would be speed exchange, and that would ultimately not produce any interaction terms at all. In the traffic example, neither energy nor momentum are conserved in interactions, but the target velocity is (with some margin of error) known in advance: The driver will attempt to go as fast as traffic permits without causing a collision. Moreover, the velocity adjustment cannot be instantaneous! Whenever we write a Boltzmann or Enskog collision term, we are implicitly assuming that "collisions" are instantaneous, and hence so is the velocity change. However, in traffic scenarios the interaction time between two cars is of comparable scale to the mean travel time between encounters, and this fact will be ignored if interactions are modeled by a "collision term" on the right hand side. In other words, every model of type $\partial_t f + v\partial_x f = C(f)$ with a (linear or nonlinear) collision term of scattering, Boltzmann or Enskog type violates the first property: the scales of interaction times are not chosen realistically. In fact, because the Enskog model compresses these times to zero, they create the option to guess a "scattering" kernel $\sigma_B(\ldots)|w - v_+|$. In the next section we will see that this option disappears if we include positive braking and acceleration times. The corresponding kinetic equation then becomes an equation of Fokker-Planck type.

4.3 "Fokker-Planck" Multilane Traffic

Focus on an individual driver on a one-lane highway, with position $x(t)$ and speed $v(t)$. x and v will change according to laws

$$\dot{x} = v, \quad \dot{v} = B(\ldots),$$

and all the detail of the interaction with traffic will be in the dependencies of the force (braking or acceleration) term $B(\ldots)$, typically augmented by a diffusive correction. If lane-changing is ignored (trivial if there is only one lane) and if there is no diffusion, the statistical dynamics of a kinetic distribution density $f(x, v, t)$ is given by a conservative transport equation

$$\partial_t f + v\partial_x f + \partial_v(B(\ldots)f) = 0$$

(this does require a derivation, but a standard one, which we omit). A diffusive correction with diffusivity $D(\ldots)$ will appear as a term $-D(\ldots)\partial_v f$ inside the last bracket.

If there are two lanes, and we have lane changing rates $p_1(\ldots)$ from lane 1 to lane 2 and $p_2(\ldots)$ from lane 2 to lane 1, then the corresponding system for the lane densities f_1 and f_2 will be

$$\partial_t f i + v\partial_x f_i \partial_v(B(\ldots)f_i - D_i(\ldots)\partial_v f_i) = p_k f_k - p_i f_i \tag{32}$$

where $k = 3 - i$.

Let us summarize: The braking/acceleration force is modeled by B, the diffusivity by D, and the lane-changing by the p_is. The details of the model are, of course, contained in the dependencies of these quantities on the state of the traffic; there, we can include such effects as nonlocalities or time-delays, driver errors leading to diffusion, and individual reaction times.

In view of what was said in the previous section the above models will assume that lane-changing is instantaneous, but braking or acceleration are not. This seems indeed reasonable; while lane-changing does in reality take some time, the process itself should be equated to the *signalling of a lane change*, because the traffic in the adjacent lane should react to that rather than the actual lane change.

Ideally, and as stated earlier, the dependencies of B and D (and p_i) should be derived from a Liouville equation via hierarchies and a closure procedure, but as such procedures are frustratingly elusive for the problem at hand, we will simply consider dependencies that seem reasonable from a heuristic point of view.

4.3.1 Dependencies.

Let us now fill in the (...) in the suggested Fokker-Planck systems. Without restricting the generality, we can scale time and distance such that $v_{min} = 0$ and $v_{max} = 1$. By H_0 we will denote a minimal safety distance (including the typical car length, as we will interpret the position of a car as the position of the *front* of the car). In addition, traffic observations tell us that there are three characteristic (reaction) times, namely, 1) the intrinsic average reaction time τ (typically of the order of 0.5 to 1.5 seconds), 2) a "braking threshold" time T_B, and 3) an "acceleration threshold" time T_A. Observations suggest that $\tau < T_B < T_A$ (the latter two being of the order of magnitude of a few seconds), and their meaning is as follows: Abbreviating T_X where $X = B$ or $X = A$, a driver moving at speed v will brake (*or accelerate*) only if the distance $H_X := H_0 + T_X v$ is smaller (*or larger*) than the distance to the leading car, and this leading car moves with a speed v_+ such that $v_+ < v$ (*or* $v_+ > v$) for $X = B$ (or $X = A$, respectively).

Drivers will be able to observe macroscopic density (ρ) and flux ($j = \rho u$), but not much additional information about the kinetic density f is directly observable (recall that f really only makes sense from a statistical point of view). Hence it was suggested (in [20]) that B, D and the $p_i's$ should only depend on these quantities, with appropriate nonlocalities. To this end, the abbreviations $\rho^B(x,t) = \rho(x + H_0 + T_B v, t - \tau)$ and $u^B(x,t) = u(x + H_0 + T_B v, t - \tau)$ (and similarly ρ^A, u^A) are conveniently introduced; note that ρ^B and u^B become dependent on the speed variable v and make no sense outside of a kinetic model.

ρ, j, u and f are naturally linked via

$$\rho = \int f \, dv, \qquad j = \int vf \, dv.$$

We set the lane change rates p_k as $p_k := P_k(\dots) j_k$, where j_k is the flux on the k-th lane and P_k is the lane change probability per car. This is dimensionally correct, although the exact dependence of p_k on j_k is debatable. In [20], the dependence of P_k on lane k was chosen as

$$P_k = P(u_k^B, v) = \begin{cases} \left(\frac{v - u_k^B}{v_{max} - u_k^B}\right)^\delta & \text{if } v > u_k^B \\ 0 & \text{if } v \le u_k^B \end{cases} \tag{33}$$

In other words, the lane changing probability is set to depend on the relative (scaled) speed with respect to the leading car; the conditions $v > u_k^B$ are implicit conditions because the right-hand side depends on v. In this sense, the definition of P_k is already quite complicated, yet we have not even included

any dependence on the state of traffic on the adjacent lane (as one should have). The number δ is a parameter, usually chosen as 1.

We next describe the ansatz for the (braking/acceleration) force B. It is

$$B_k = B[f_k](x,v,t) = \begin{cases} -c_B(v-u_k^B)^2\rho_k^B(1-P(u_k^B,v)) & \text{if } v > u_k^B \\ c_A(v-u_k^A)^2(\rho_{max}-\rho_k^A) & \text{if } v \le u_k^B \text{ and } v \le u_k^A \\ 0 & \text{otherwise} \end{cases} \tag{34}$$

Several remarks are in order. First, the chosen powers of the relative speeds $(v-u)$ and the densities allow for dimensionless constants c_B, c_A. Other powers (in particular the first power for the relative speed) are certainly worthy of attention. Second, note that the braking force is chosen to be proportional to ρ for braking scenarios, but proportional to $\rho_{max}-\rho$ for acceleration conditions. This seems reasonable, as higher densities make acceleration less likely. In any case, these dependencies are nothing but guesses, and one could substitute a general function $g(\rho)$. Third, the reader may wonder why there are three lines in (34). This is because it is conceivable that the first two cases do not cover all possibilities, because u^A and u^B are not computed at the same spot. Imagine a situation where you move slower than the traffic immediately in front (at $x+H_0+T_Bv$), but faster than the traffic further ahead (at $x+H_0+T_Av$). What would you do? We suggest you should do nothing. Fourth, and last, the braking probability $(1-P)$ is only present in braking scenarios.

We mention that if we assume that $T_B = T_A =: T$, a much simpler braking ansatz with the same basic qualitative features is

$$B = B[f](x,v,t) = -c(v-u^B)g(\rho^B). \tag{35}$$

Note that we will brake if $v > u(x+H_0+Tv)$ and accelerate otherwise, and we have left the $\rho-$ dependence open. It has recently been observed ([19]) that the Aw-Rascle model can be derived from moment equations associated with this model.

Finally, we need a reasonable idea for the diffusion. In [20] the ansatz

$$D_k(\ldots) = \sigma(\rho_k,u_k)|v-u^B|^\gamma \tag{36}$$

was chosen, with $\gamma \in (1,3)$. The function σ was set up to vanish fast enough for the limiting values of ρ and u (0 and ρ_{max}, v_{max}). Clearly, these limiting values are of little significance for practical applications, as the behaviour of the cars on the road is a priori known at these values. We will therefore not discuss σ at all. The values of γ and the degeneracy of the diffusion

at $v = u^B$ were chosen intentionally: with the given form, this diffusion is consistent with the existence of the trivial (synchronized) equilibria $\rho\delta_u(v)$. The verification of this fact is an easy exercise in distribution theory.

We pause to revisit our list of desirable properties of a kinetic traffic model. It should be transparent at this point that the Fokker-Planck models as described so far satisfy the first two properties: realistic scales and existence of synchronized equilibria. We will shortly see that fundamental diagrams are computable (so the third item on our list is covered), and it has been shown in numerical experiments that relaxation to identical equilibria on both lanes, as well as the formation of stop-and-go waves behind bottlenecks are predicted by these models. So the only significant gap in the program is the last item, namely, a satisfactory derivation of the model from a Liouville equation.

The diffusion D_k as given above vanishes at $v = u^B$, a deliberate choice. Is that realistic? It means that if a driver moves at exactly the speed he/she sees in the lead traffic, he/she will not adjust his/her speed at all, not even by a nervous foot on the gas pedal. This seems somewhat beyond human precision, though it appears that in reasonable high density regimes drivers will concentrate to such an extent that the errors become very small. In general, though, and in particular in lighter traffic, when there is more distance between cars, some of the drivers will be prone to speed fluctuations due to their judgement (or lack thereof). A more realistic diffusion function should therefore be of the type

$$D_k(\ldots) = \sigma(\rho_k, u_k)|v - u^B|^\gamma + \epsilon(\rho, u).$$

This will remove the synchronized equilibria in all regimes where $\epsilon(\rho, u) > 0$. One can speculate about the dependencies of this "residual diffusion"; as a first approximation we suggest to take ϵ as a (small) positive constant.

This completes our model description.

4.4 Spatial and lane homogeneity, non-trivial equilibria, and computing fundamental diagrams

There are several possibilities of homogeneous scenarios: identical, but time- and space- dependent traffic on all lanes (lane homogeneity), or spatial homogeneity (i.e., the densities for all lanes are independent of x), or both. In the last case the equations simplify dramatically, as the lane-changing terms will

cancel and the nonlocalities become irrelevant. We are left with a nonlinear drift-diffusion equation

$$\partial_t f + \partial_v(B([f], v)f - D([f], v)\partial_v f) = 0$$

where $f = f_i, i = 1, 2$ (identical densities on either lane), and B and D depend on v and the moments of f as described in the previous section. We have emphasized this dependence in the above equation but will in the sequel just write $B(\ldots)$, etc.

The equation should be complemented with zero flux (Robin) boundary conditions at $v = 0$ and $v = v_{max}$:

$$B(\ldots)f = D(\ldots)\partial_v f$$

at $v = 0, v_{max}$. Equilibria are time-independent, space- and lane- homogeneous solutions of the model, which also have to satisfy these Robin boundary conditions. This means they have to satisfy the nonlinear first order ordinary differential equation

$$B(\ldots)f = D(\ldots)\partial_v f. \tag{37}$$

The nonlinearity is, of course, given via the dependence of B and D on the macroscopic density and flux. These dependencies introduce a rather unusual nonlocality into the coefficients of the equation.

If D is degenerate (i.e., D vanishes at $v = u$) then Eq. (37) applies only for $v \neq u$. This causes several problems:

- First, for the dependencies given in (34) the braking (or acceleration) force also vanishes at $v = u$. The implication is that there will be no flux through the average speed, and this suggests that relaxation to equilibria (to be discussed later) will occur for $v > u(t)$ and $v < u(t)$, even while $u(t)$ evolves with f, but the equilibria values for the two domains may not link continuously at u. This is indeed the case and has been seen in numerical experiments (see [16]). Even if we start with a very smooth f, a jump will develop in the limit $t \to \infty$.

- These discontinuous equilibria were ignored in [20]; there, only continuous solutions of (37) were allowed.

- On the upside, when the degeneracy is strong enough ($\gamma > 1$) it permits the trivial equilibria $\rho\delta(v-u)$. We realize that it is our desire to include these trivial equilibria that introduces, so to speak via the back door, additional equilibria which may not be of real interest.

As shown in Ref. [20], for any given ρ and u, continuous solutions for the degenerate case and the B as given in (34) can be computed explicitly by integration of (37). Only the (rather pathological) boundary points $\rho = 0, u = v_{max}$ and $\rho = \rho_{max}, u = 0$ need some special attention (this is where the function σ in (36) comes in), but we will not discuss this here. The solutions then need to be normalized such that $\int f\ dv = \rho$, an easy step, and this leaves us then with a mapping $u \to f[u]$ (because the equation depends parametrically on u, so do the normalized solutions). The *fundamental diagram* is defined as the set of all fixed points of this mapping. In other words, once ρ is chosen in the viable domain, we have to look for an u such that $\rho u = \int v f[u](v) dv$, or equivalently, u has to be a root of

$$R_\rho(u) := \int_0^{v_{max}} (v - u) f[u](v)\ dv.$$

It requires little effort to see that $R_\rho(u) > 0$ for u near 0, and $R_\rho(u) < 0$ for u near v_{max}, and hence continuity implies that there is at least one root. Hence the fundamental diagram is well defined.

This discussion summarizes the analysis given in [20]. In that paper it was shown that if lane change probabilities of the type given in (33) are included, then there is an interval (ρ_1, ρ_2) of densities such that the fundamental diagram is three-valued for $\rho \in (\rho_1, \rho_2)$ (and only there). This fundamental diagram is depicted in Figure 6.

Several questions arise in this context, and most of these questions have been answered (some of the answers have not appeared in print, but they hopefully will in due time):

1. What happens if the degeneracy is removed? Remember that there is a good rationale for this– the uncertainty of the individual driver to read or respond accurately to a traffic scenario ahead. Think of a residual diffusion $\epsilon > 0$ to be added to D. It should be clear that this eliminates the trivial synchronized equilibria; it is also clear that it removes the piecewise continuous equilibria with a jump at u, simply because the defining differential equations can now be integrated past this point. While the solvability of (37) for given ρ and u in this case is clear, it is much harder (though possible, [19]) to compute explicit formulas for these equilibria (MAPLE helps). The fundamental diagram continues to be multi-valued while the residual diffusion is small enough.

2. We state that it is the lane-changing that causes the multi-valued fundamental diagram. Indeed, suppose we turn off the lane-changing (a

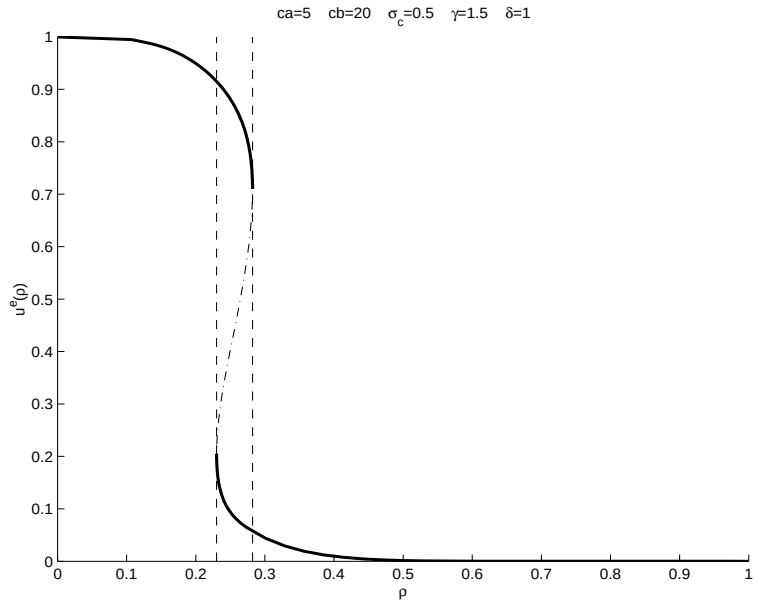

Figure 6: A multivalued fundamental diagram

The lane-changing probability leads to a density regime where several equilibria with different average speeds are possible.

convenient way to do this in (33) is to set $\delta = \infty$). It turns out that then for every $\rho \in [0, \rho_{max}]$ the function

$$u \to R_\rho(u)$$

is strictly monotone increasing, and will therefore possess exactly one root. In fact, this property depends only on the monotonicity of $\frac{B}{D}$ as a function of u (see [36] for an example), and it is exactly the presence of a lane-changing probability that may alter this monotonicity in some domains.

3. If we are in the region where there is more than one equilibrium, which of these are stable? While it is easy to make guesses on this ("the two with high and low flux are stable, the one associated with the middle branch and intermediate flux is not") it turns out to be a nontrivial matter to decide. Analytical and numerical work on this is in progress.

4. A simpler question of practical importance concerns the number of lane-changes associated with an equilibrium. If f_e is the equilibrium associated with the point on the fundamental diagram (ρ, u), the average number of lane changes for traffic in this equilibrium is

$$\int P(u,v) f_e(v)\ dv.$$

Using the definition of P given in (33) and the equilibria computed in [20], one easily sees that there is more lane changing in the equilibrium associated with the higher flux. This is intuitive but sounds dangerous; however, speaking realistically, it is only in moderate densities that multi-valued fundamental diagrams are expected; in higher densities, lane-changing becomes difficult (and more dangerous), and the model should be adjusted for this. That lane-changing should intrinsically be a positive factor towards flux enhancement should be intuitively clear.

4.5 Further Results

We conclude with a brief synopsis of other known facts on the Fokker-Planck models. In situations where both lanes are treated equally and lane-changing is symmetric, it is to be expected that traffic synchronizes in the sense that nearly identical distributions should emerge on both lanes after a short time (note that this concept of synchronization is different from the one discussed earlier, in which all vehicles assume an identical speed). Some numerical experiments showing that such synchronization holds for the FP models were given in [16]; some analytical result for simplified equations can also be found in that reference.

Other analytical work, done in [5], concerns the relaxation of space- and lane-homogeneous solutions to one of the equilibria computed for the setup of the fundamental diagram. This is a question which invites the use of entropy methods such as widely used in the literature on PDEs such as reaction-diffusion, porous media or Fokker-Planck type equations (see for example [3]). The basic idea is to study the long-term behavior of relative entropy functionals of the type

$$E[f|g](t) = \int \Phi\left(\frac{f}{g}\right)\ dv.$$

Here, Φ is a suitably chosen convex function (for example, $\Phi(x) = (x^p - x)/(p-1)$ for some $p > 1$), $f = f(t,v)$ is the spatially homogeneous traffic density, and $g = g(t,v)$ is a local equilibrium, i.e. a function for which the drift-diffusion part of the kinetic model vanish; but g is in general no solution because it depends on time—only if g is associated with a density-flux pair on the fundamental diagram is it a steady solution. Standard calculations for such entropy functionals and their associated entropy productions may be employed to obtain results on the asymptotic convergence of $f(t,\cdot)$ to a

steady equilibrium. However, this is delicate because for the model under consideration the local equilibria depend on the first moment $u(t)\rho(t)$ of f, and we know that at least for some values of ρ there are several steady equilibria. It is therefore not clear what f will do as $t \to \infty$; the result in [5] states that convergence to a steady solution is to be expected, but we could not state to which one, and we could not find convergence rates.

It is reasonable to expect that realistic traffic models should pose profound difficulties from an analytical point of view; after all, they should also predict rather complex behaviour. The presented Fokker-Planck type models do both, but there is no doubt in my mind that they are still far from reality. The truth, as always, is stranger.

I would like to end this article with words of thanks to the organizers of the Porto Ercole summer schools and to the students and colleagues who listened to my lectures. Thank you for your interest.

References

[1] B.Aw, A. Klar, Th. Materne, M. Rascle (2002). *Derivation of continuum traffic models from microscopic follow-the-leader models.* SIAM J. Appl. Math. **63/1**, 259-278

[2] O.G. Berg, R.B. Winter, and P.H. von Hippel (1981). *Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory.* Biochemistry **20**, 6929-6948

[3] J. A. Carrillo, A. Jüngel, P. A. Markowich, G. Toscani, and A. Unterreiter (2001). *Entropy dissipation methods for degenerate parabolic problems and generalized Sobolev inequalities.* Monatsh. Math. **133**, 1-82

[4] C.F. Daganzo (1995). *Requiem for second order fluid approximations of traffic flow.* Transportation Research B, **29 B**, 277-286

[5] J. Dolbeault, R. Illner (2003). *Entropy methods for kinetic models of traffic flow.* Comm. Math. Sci. **1**(3), 409-421

[6] H.B. Dowse, J.M. Ringo (1987). *Further evidence that the circadian clock in Drosophila is a population of coupled ultradian oscillators.* Journal of Biological Rhythms **2:1**, 65-76

[7] J.C. Dunlap (1999). *Molecular bases for circadian clocks.* Cell **96**, 271-290

[8] R. Edwards, R. Gibson, R. Illner, and V. Paetkau (2007). *Coupled stochastic ordinary differential equations and circadian rhythms.* Theoretical Biology and Medical Modelling, to appear

[9] R. Edwards, R. Illner, and V. Paetkau (2006). *A model for generating circadian rhythm by coupling ultradian oscillators.* Theoretical Biology and Medical Modelling **3:12 (23 Feb 2006),** http://www.tbiomed.com/content/3/1/12

[10] M.B. Elowitz, S. Leibler (2000). *A synthetic oscillatory network of transciptional regulators.* Nature **403**, 335-338

[11] B. Ermentrout (2002). *Simulating, analyzing, and animating dynamical systems. A guide to XPPAUT for researchers and students.* SIAM Publ.

[12] D.T. Gillespie (1977). *Exact stochastic simulation of coupled chemical-reactions.* J. Phys. Chem. **81**, 2340-2361

[13] D. Gonze, J. Halloy, J.C. Leloup, and A. Goldbeter (2003). *Stochastic models for circadian rhythms: effect of molecular noise on periodic and chapotic behaviour.* C.R. Biol **326**, 189-203

[14] T.E. Gough, R. Illner (1998). *Modeling the solid state reaction* $CO_2.C_2H_2 \to CO_2 + C_2H_2$. Chem. Phys. Lett.298, 196-200

[15] T.E. Gough, R. Illner (1999). *Modeling crystallization dynamics when the Avrami model fails.* VLSI Design **9**(4), 377-383

[16] M.Herty, R. Illner, A. Klar, and V. Panferov (2006). *Qualitative properties of solutions to systems of Fokker–Planck models for multilane traffic flow.* Transport Th. Stat. Phys., to appear

[17] M. Herty, M. Rascle (2005). *Coupling conditions for a class of second order models for traffic flow.* SIAM J. Math. Anal. **38**(2), 155-169

[18] R. Illner, S. Bohun, S.McCollum, and Th. van Roode (2005). *Mathematical Modelling: A Case Studies Approach.* AMS Student Library vol. 27

[19] R. Illner, C. Kirchner, and R. Pinnau (2007). *A derivation of the Aw-Rascle traffic models from Fokker-Planck type kinetic models.* preprint

[20] R. Illner, A. Klar, and Th. Materne (2003). *Vlasov-Fokker-Planck Models for Multilane Traffic Flow.* Comm. Math. Sci. **1**, 1-12

[21] R. Illner, A. Klar, H. Lange, A. Unterreiter, and R. Wegener (1999). *A kinetic model for vehicular traffic: existence of stationary solutions.* J. Math. Anal. Appl. **237**, 622-643

[22] R. Illner, A. Klar, C. Stoica, and R. Wegener (2002). *Kinetic equilibria in traffic flow models.* Transport Th. Stat. Phys. **31**(7), 615-634

[23] B.S. Kerner (1998). *Experimental features of self-organization in traffic flow.* Phys. Rev. Letters **81**, 3797-3800

[24] B.S. Kerner (2000). *Experimental features of the emergence of moving traffic jams in free traffic flow.* J. Phys. A **33**, 221-228

[25] A. Klar, R. Wegener (1998). *A hierarchy of models for multilane vehicular traffic II: Numerical Investigation.* SIAM J. Appl. Math. **59**, 983-1002

[26] A. Klar, R. Wegener (1998). *A hierarchy of models for multilane vehicular traffic I: Modeling.* SIAM J. Appl. Math. **59**, 1002-1011

[27] I. Mihalcescu, W. Hsing, and S. Leibler (2004). *Resilient circadian oscillator revealed in individual cyanobacteria.* Nature **430**, 81-85

[28] J.C. Leloup, A. Goldbeter (2003). *Toward a detailed computational model for the mammalian circadian clock.* Proc. Natl. Acad. Sci. USA **100**, 7051-7056

[29] A.D. Riggs, S. Bourgeois, and M. Cohn (1970). *The lac repressor-operator interaction. 3. Kinetic studies.* J. Mol. Biol. **53**, 401-417

[30] U. Schibler, F. Naef (2005). *Cellular oscillators: rhythmic gene expression and metabolism.* Curr. Opin. Cell Biol. **53**, 401-417

[31] H. Struchtrup (2005). *Macroscopic Transport Equations for Rarefied Gas Flows.* Springer-Verlag

[32] S.M. Ross (1989). *Introduction to Probability Models.* 4th Ed., Academic Press

[33] J.M. Vilar, H.Y. Kueh, N. Barkai, and S. Leibler (2002). *Mechanisms of noise-resistance in genetic oscillators.* Proc. Natl. Acad. Sci. USA **99**, 5988-5992

[34] A.T. Winfree (1975). *Unclocklike behaviour of biological clocks.* Nature **253**, 315-319

[35] R.B. Winter, O.G. Berg, and P.H. von Hippel (1981). *Diffusion-driven mechanisms of protein-translocation on nucleic acids. 3. The Escherichia coli lac repressor-operator intercation: kinetic measurements and conclusions.* Biochemistry **20**, 6961-6977

[36] T. Zhou (2006). *Vlasov-Fokker-Planck type kinetic models for multilane traffic flow, and large-time behaviour of the kinetic density by entropy methods.* University of Victoria, M.Sc. thesis