

Stochastic Dynamics in Systems Biology
Lecture Notes

V. Wolf¹ P. Crouzen²

July 27, 2010

¹wolf@cs.uni-sb.de

²crouzen@cs.uni-sb.de

Contents

1	Deterministic vs. Stochastic Chemical Kinetics	1
1.1	Introduction	1
1.2	Reaction Rate Equations	2
1.2.1	Specification of Reaction Networks	2
1.2.2	Law of Mass Action	3
1.2.3	Michaelis-Menten Approximation	8
1.3	Stochastic Phenomena in the Cell	9
1.3.1	Lambda Phage Decision Circuit	9
1.3.2	Intrinsic and Extrinsic Molecular Noise	11
1.4	Basic Concepts of Probability	12
1.4.1	Probabilities	12
1.4.2	Discrete Random Variables	15
1.4.3	Continuous Random Variables	22
1.5	Stochastic Chemical Kinetics	26
1.5.1	Transition Probabilities	28
1.5.2	Chemical Master Equation	31
1.5.3	Expected Populations	33
1.6	Summary	36
2	Stochastic Simulation	37
2.1	Introduction	37
2.2	Trajectory Generation	37
2.2.1	Residence Times	38
2.2.2	Number of Jumps	39
2.2.3	Jump Probabilities	40
2.2.4	Algorithm	42
2.3	Output Analysis	44
2.3.1	Fundamental Results from Statistics	44
2.3.2	Interval Estimation	47
2.4	Summary	50

3	Numerical Solution Methods	53
3.1	Introduction	53
3.2	Matrix-Vector Form and Matrix Exponential	54
3.3	Uniformization	55
3.4	Summary	59

Preface

These lectures were delivered at Saarland University in 2009. They focus on the analysis of stochastic models in systems biology. Knowledge of basic mathematical concepts is a prerequisite for understanding these lecture notes. The necessary background in systems biology and probability theory is part of these notes.

We use hyperlinks to Wikipedia pages to encourage students to explore further the context of this course. However, we would like to point out that Wikipedia entries are mostly written by laymen and should not be used as a primary information source.

Chapter 1

Deterministic vs. Stochastic Chemical Kinetics

1.1 Introduction

Mathematical models are an indispensable part of the systems biology research cycle. They are used to complement experimental studies that are carried out *in vivo* or *in vitro* in order to refine, falsify and verify hypotheses related to biological phenomena. The *in silico* simulation of a biological system based on a mathematical model is fast and does not require expensive laboratory work. Moreover, it can reveal details about the functional relationships between proteins and other molecules in the cell.

The traditional modeling approach for dynamic systems of cellular processes is based on *deterministic* models, where the future of the system can be predicted with certainty. They describe the interactions between populations of molecules of different types. Instead of modeling each single molecule in detail (position in space, functional structure, etc.), this modeling approach operates at a macroscopic scale. The state of the system is given by the concentrations of the chemical species and a continuous deterministic change of these concentrations is assumed.

Here we focus on *stochastic* models. They are based on the assumption that the state of the system changes at discrete points in time (see Fig. 1.1). These changes are triggered by chemical reactions that occur randomly. The stochastic approach has gained more and more attention recently since in many biological systems the randomness of microscopic events have a significant influence. In a stochastic model,

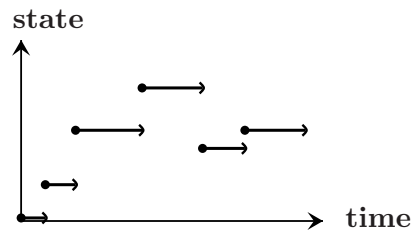


Figure 1.1: State changes occur at discrete points in time.

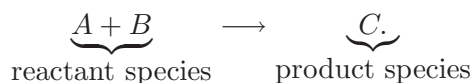
various possibilities exist for the future behavior of the system, where each possibility has a certain probability. In this chapter, we motivate and discuss the definition of discrete-state stochastic models for dynamic cellular processes and compare them to deterministic models.

1.2 Reaction Rate Equations

Reaction rate equations are a deterministic modeling approach that is the most widely used approach for cellular dynamics. They are based on the idea that the concentrations of the chemical substances can be approximated by a process that changes *continuously* and *deterministically* in time.

1.2.1 Specification of Reaction Networks

Consider the *stoichiometric equation*



The uppercase letters denote different types of molecules (also called *chemical species*). We refer to the chemical species on the left hand side of the arrow as *reactant species* and to those on the right hand side as *product species*. A stoichiometric equation describes a chemical reaction and since several instances of the same reaction may occur, it is also referred to as a *reaction channel*. Stoichiometric equations specify which reactant species are required for the reaction to occur and which are the products of the reaction. In the above example, a molecule of type A and a molecule of type B form a molecule of type C . Note that reactions may require/produce more than one molecule of a certain type. In this case a *stoichiometric coefficient* is explicitly added, e.g. consider the reaction channel



which describes a dimerization.

In the sequel, we consider only *elementary reactions* which correspond to a single mechanistic step. In general, reactions may have intermediate products and/or parallel reaction pathways. They can, however, always be decomposed into elementary reactions. We assume that the number of reactant and product species is at most two, since it is highly unlikely that more than two molecules collide at the same time.

Mathematical models of natural systems are usually idealized descriptions of the real system since

- the real system is extremely complex and a detailed description would make the model intractable,

- many details have no significant influence on the phenomenon the modeler wants to study with the mathematical model.

The following assumptions for mathematical models of coupled chemical reactions simplify the construction and analysis of the model considerably and have proven to be appropriate for most systems.

We consider a

- constant reaction volume
- with fixed temperature and pressure
- that is spatially homogeneous (i.e. a well-stirred mixture).

In this section, $A(t)$ denotes the molar concentration of type A molecules at time instant $t \geq 0$, that is, the number of moles of A per liter, where one mole contains N_A molecules. The number $N_A \approx 6 \cdot 10^{23}$ is called the *Avogadro constant*. Clearly, the total number $\#A(t)$ of molecules of type A at time t is

$$\#A(t) = A(t) \cdot N_A \cdot V,$$

where V is the volume (in liters).

1.2.2 Law of Mass Action

Reaction rate equations are based on the law of mass action, which reasons about the change of the chemical concentrations. According to the law of mass action, for a small time step of length $\Delta > 0$, the change

$$C(t + \Delta) - C(t)$$

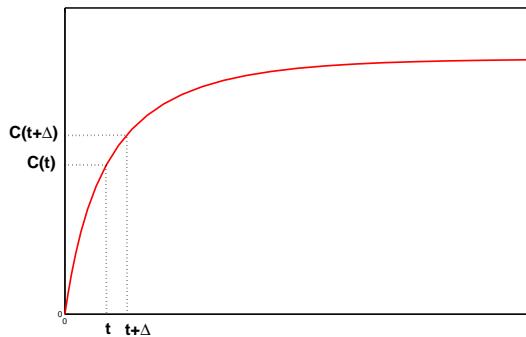
of the concentration of a product C of a reaction is

proportional to the product of the reactant concentrations and Δ .

For instance, in the case of $A + B \rightarrow C$ we have

$$C(t + \Delta) - C(t) \approx k \cdot A(t) \cdot B(t) \cdot \Delta.$$

for small Δ . This approximation becomes exact as Δ approaches 0.



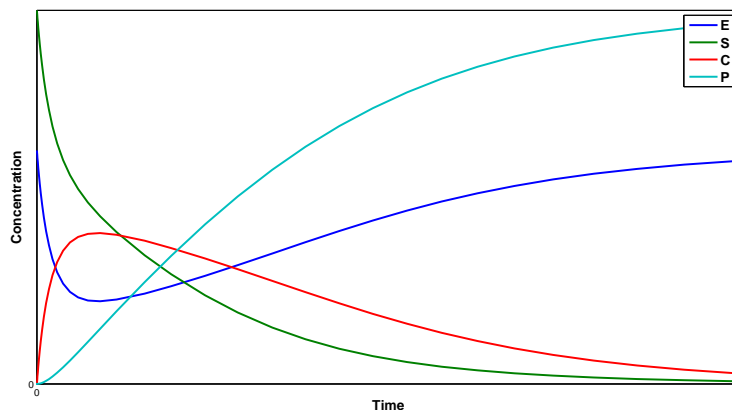


Figure 1.3: ODE solution of the enzyme reaction.

If we know the concentration of C at time t we can approximate $C(t + \Delta)$ by $C(t) + k \cdot A(t) \cdot B(t) \cdot \Delta$. The factor $k > 0$ is called reaction rate constant or *affinity constant*. Letting $\Delta \rightarrow 0$, we obtain

$$\lim_{\Delta \rightarrow 0} \frac{C(t+\Delta) - C(t)}{\Delta} = \frac{d}{dt} C(t) = k \cdot A(t) \cdot B(t). \quad (1.1)$$

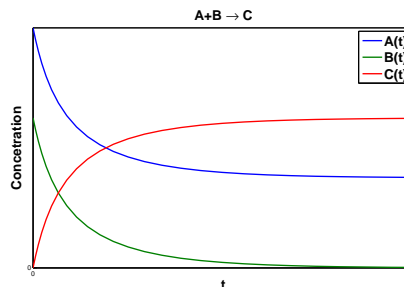


Figure 1.2: Solution of the system of differential equations for $A + B \rightarrow C$.

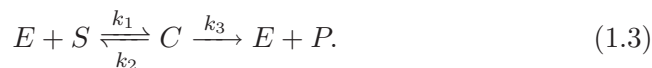
The remaining species A and B are used up by the reaction $A + B \rightarrow C$ and thus

$$\frac{d}{dt} A(t) = \frac{d}{dt} B(t) = -k \cdot A(t) \cdot B(t). \quad (1.2)$$

The system of differential equations given by (1.1) and (1.2) is called *reaction rate equations*. Note that we have to specify initial concentration (that is, values for $A(0), B(0), C(0)$) in order to obtain solutions $A(t), B(t), C(t)$ for all $t \geq 0$.

Example 1: ENZYME KINETICS

We consider a network of three biochemical reactions given by



It describes the formation of a complex (C) from an enzyme (E) and a substrate (S). The complex molecule can either dissociate to yield E and S

again or to yield E and a product P . The corresponding differential equations for the concentrations are

$$\begin{aligned}\frac{d}{dt}E(t) &= -k_1 \cdot E(t) \cdot S(t) + (k_2 + k_3) \cdot C(t) \\ \frac{d}{dt}S(t) &= -k_1 \cdot E(t) \cdot S(t) + k_2 \cdot C(t) \\ \frac{d}{dt}C(t) &= k_1 \cdot E(t) \cdot S(t) - (k_2 + k_3) \cdot C(t) \\ \frac{d}{dt}P(t) &= k_3 \cdot C(t).\end{aligned}$$

Note that since this system is closed (molecules can neither “disappear” nor “appear from nothing”), we have the following conservation laws. Assume $E(0) = e_0$, $S(0) = s_0$, and $C(0) = P(0) = 0$. Then for all t ¹

$$\begin{aligned}E(t) + C(t) &= e_0 \\ S(t) + C(t) + P(t) &= s_0\end{aligned}$$

This can be easily seen by checking

$$\begin{aligned}\frac{d}{dt}E(t) + \frac{d}{dt}C(t) &= 0 \\ \frac{d}{dt}S(t) + \frac{d}{dt}C(t) + \frac{d}{dt}P(t) &= 0.\end{aligned}$$

Fig. 1.3 shows the solution of the above system of differential equations. Initially the concentrations of P and C are zero. Then $P(t)$ increases until all molecules of type S are transformed into P molecules. Thus, $S(t)$ decreases in time. The concentration of complex molecules raises quickly at the beginning but decreases afterwards since the number of substrates becomes exhausted.

In general, the reaction rate equations yield a system of *non-linear* ordinary differential equations (ODEs). A general form for a system of (coupled) ODEs is

$$\frac{d}{dt}y(t) = f(t, y(t)),$$

where $t \in \mathbb{R}$, $y : \mathbb{R} \rightarrow \mathbb{R}^n$, $f : \mathbb{R}^{n+1} \rightarrow \mathbb{R}^n$. In the case of reaction rate equations, n is the number of chemical species, $y(t)$ is a vector of concentrations and $f(t, y(t))$ is a vector whose entries specify the change of the concentrations.

Solving Ordinary Differential Equations We give a short introduction on the solution of ordinary differential equations. For a more detailed discussion we refer to [HNW08, HW04].

We consider the following general form of a system of ODEs

$$\frac{d}{dt}y(t) = f(t, y(t)), \tag{1.4}$$

¹We assume $C(0) = P(0) = 0$ for simplicity. The case $C(0) = P(0) > 0$ yields similar but slightly more complex conservation laws, which we leave as an exercise.

where $t \in \mathbb{R}$, $y : \mathbb{R} \rightarrow \mathbb{R}^n$, and $f : \mathbb{R}^{n+1} \rightarrow \mathbb{R}^n$. Eq. (1.4) specifies the slope of the unknown function $y(t)$ but not the actual values. In general, there is an infinite family of solutions if f is “sufficiently smooth” (more precisely, f is Lipschitz continuous in y and continuous in t). If a system of ODEs is derived from reaction rate equations, the smoothness condition always holds for bounded time intervals. Informally, this comes from the simple product form of the law of mass action. If a single value $y(t_0) = y_0$ is specified at some point t_0 , there is a unique function $y(t)$ with $y(t_0) = y_0$ that satisfies Eq. (1.4). It is called the *solution* to the initial value problem defined by Eq. (1.4) and the point $y(t_0) = y_0$.

In simple cases, it is possible to derive an analytical expression for y .

Example 2: EXPONENTIAL FUNCTION

Let $\lambda \in \mathbb{R}$. The one-dimensional ODE

$$\frac{d}{dt}y(t) = \lambda \cdot y(t), \tag{1.5}$$

has solutions $y(t) = c \cdot e^{\lambda t}$ for $c \in \mathbb{R}$ (compare Fig. 1.4). If we fix $y(0) = y_0$ then $y(t) = y_0 \cdot e^{\lambda t}$.

Informally, an ODE is called *unstable* if the members of the solution family move away from each other with time (see Fig. 1.4, top) and *stable* if they move closer. If neither of these two cases are true, the ODE is called *neutrally stable*. An example for a neutrally stable ODE is $\frac{d}{dt}y(t) = a$ for $a \in \mathbb{R}$, which has the solution $y(t) = a \cdot t + c$.

Note that stable or unstable behavior can occur in different parts of the domain for the same equation. In general, for $t \in \mathbb{R}$, the Jacobian matrix

$$\{J\}_{ij} = \frac{df_i(t, y(t))}{dy_j(t)}$$

provides information about the stability of the ODE. If there is any eigenvalue with positive real part, the ODE is unstable at t . If all eigenvalues have negative real part, the ODE is stable at t . For instance, the Jacobian matrix of the ODE in Eq. (1.5) is $J = \lambda$ (and thus the only eigenvalue is λ).

Let us now consider the approximation of the ODE solution at discrete points. The idea is to start with the initial condition $y(t_0) = c$ and predict $y(t_0 + \Delta)$ for a small increment Δ .

As a prototype approach we discuss the Euler method. Most other approximation methods for ODEs are based on similar ideas. From

$$\lim_{\Delta \rightarrow 0} \frac{y(t + \Delta) - y(t)}{\Delta} = \frac{d}{dt}y(t) = f(t, y(t))$$

it follows that, for small $\Delta > 0$,

$$y(t + \Delta) \approx y(t) + f(t, y(t)) \cdot \Delta.$$

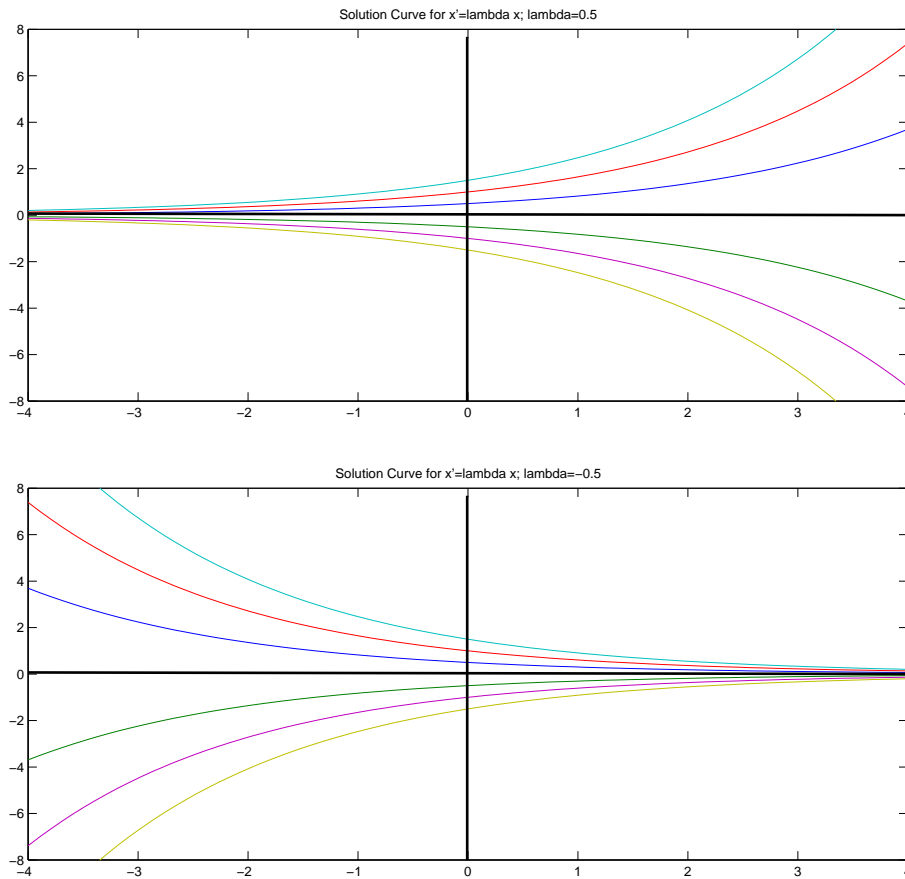


Figure 1.4: Solution curves for $\frac{d}{dt}y(t) = 0.5 \cdot y(t)$ (top) and $\frac{d}{dt}y(t) = -0.5 \cdot y(t)$ (bottom).

For initial value c this leads to the iteration

$$\begin{aligned} t_0 &= 0, & t_{n+1} &= t_n + \Delta \text{ for } n \geq 1, \\ y_0 &= c, & y_{n+1} &= y_n + \Delta \cdot f(t_n, y_n). \end{aligned}$$

The values y_n converge to the true solution $y(t_n)$ if $\Delta \rightarrow 0$. In practice, the Euler method yields poor results compared to other methods, such as Runge-Kutta methods. The reason is that in order to bound the approximation error, for many systems the stepsize has to be chosen very small. Especially, small errors are magnified with time if the equation is unstable, since the n -th approximation lies on a different member of the solution family. Errors are, however, diminished with time if the equation is stable and the step size is “small enough”.

In general, the quality of the approximation depends on the particular ODE, the chosen numerical method, and the stepsize Δ .

1.2.3 Michaelis-Menten Approximation

Solving large reaction rate equations can be computationally difficult. In this case it is useful to find a smaller ODE which has approximately the same solution as the large system of reaction rate equations. Michaelis-Menten approximation techniques can be used when some of the reactions are much faster than others. We then find that the fast reactions stabilize very quickly compared to the slow reactions. Using this information we can approximate the ODE system by assuming that the fast reactions are *always* in their stable state.

Example 3: MICHAELIS-MENTEN APPROXIMATION

Consider the enzyme reaction in (1.3). Now assume that the creation of the complex C and the degradation of the complex into enzyme E and substrate S occurs much faster than the degradation of the complex into enzyme and product. Because of this we may assume that the concentration of C is always stable, i.e., $\frac{d}{dt}C(t) = 0$. From the conservation laws we also have $E(t) + C(t) = e_0$. We now rewrite the ODE for the complex to express $C(t)$ in terms of $S(t)$:

$$\begin{aligned} 0 &= \frac{d}{dt}C(t) = k_1 \cdot E(t) \cdot S(t) - (k_2 + k_3) \cdot C(t) \\ &= k_1 \cdot (e_0 - C(t)) \cdot S(t) - (k_2 + k_3) \cdot C(t) \\ &= k_1 \cdot e_0 \cdot S(t) - (k_1 \cdot S(t) + k_2 + k_3) \cdot C(t) \\ \Leftrightarrow C(t) &= e_0 \cdot \frac{k_1 \cdot S(t)}{k_1 \cdot S(t) + k_2 + k_3} \\ C(t) &= e_0 \cdot \frac{S(t)}{S(t) + k_m}. \end{aligned}$$

Here $k_m = \frac{k_2 + k_3}{k_1}$ is called the Michaelis constant. It is important to realize that we do not assume that $C(t)$ is constant even though we assume that $\frac{d}{dt}C(t) = 0$. Rather we assume that the concentration of C changes so fast in comparison to the slow production of P that, when the concentration of S changes, the concentration of C immediately moves to its stable concentration.

Now that we have expressed $C(t)$ in terms of $S(t)$ we can also express $\frac{d}{dt}P(t)$ in terms of $S(t)$:

$$\begin{aligned} \frac{d}{dt}P(t) &= k_3 \cdot C(t) \\ \frac{d}{dt}P(t) &= k_3 \cdot e_0 \cdot \frac{S(t)}{S(t) + k_m} \end{aligned}$$

The constant k_m is a measure of how much slower the production of P (reaction rate constant k_3) is compared to the other two reactions (constants k_1 and k_2). We compare this constant with the substrate concentration S and consider three cases.

- If $S(t) \gg k_m$ then there is a large amount of substrate. This means that most enzymes will be part of a complex molecule. We find $\frac{S(t)}{S(t) + k_m} \approx$

1 and thus $\frac{d}{dt}P(t) \approx k_3 \cdot e_0$. In other words, protein production is approximately constant.

- If $S(t) \ll k_m$ then there is a small amount of substrate, but the first two reactions are much faster than the protein production reaction. Most enzymes will then be free (not part of a complex). We find $\frac{S(t)}{S(t)+k_m} \approx \frac{S(t)}{k_m}$ and thus $\frac{d}{dt}P(t) \approx \frac{k_3}{k_m} \cdot e_0 \cdot S(t)$. In other words, protein production can be seen as a reaction $S \rightarrow P$ with rate constant $\frac{k_3}{k_m} \cdot e_0$.
- If $S(t) \approx k_m$, we find $\frac{S(t)}{S(t)+k_m} \approx \frac{1}{2}$ and $\frac{d}{dt}P(t) \approx \frac{1}{2} \cdot k_3 \cdot e_0$. Again, we find that protein production is approximately constant.

1.3 Stochastic Phenomena in the Cell

“BIOLOGISTS TEND TO THINK DETERMINISTICALLY. A CASE IN POINT IS THEIR PROLONGED SEARCH FOR THE “FOUNDER CELLS” OF THE SLIME MOULD DICTYOSTELIUM. THESE AMOEBAE EMIT PULSATILE cAMP SIGNALS UNDER STARVATION CONDITIONS, MOBILIZING NEIGHBORING CELLS TO SURROUND THEM AND FORM A MOTILE MULTICELLULAR SLUG THAT WANDERS OFF TO FORM SPORES. DESPITE THEIR EFFORTS, BIOLOGISTS NEVER COULD FIND THE FOUNDER CELLS. THE REASON IS THAT ALL DICTYOSTELIUM AMOEBAE HAVE THE CAPACITY TO PRODUCE cAMP SIGNALS, AND BECOMING A FOUNDER CELL IS A MATTER OF CHANCE—IT’S A STOCHASTIC PROCESS (1).”
(In “Small Numbers of Big Molecules”)

In this section, we motivate stochastic modelling approaches for dynamical systems in the area of systems biology. We present one of the most prominent examples, the phage lambda decision circuit, where a deterministic model (such as the reaction rate equations introduced in the previous section) is inappropriate and a stochastic model is required [ARM98].

1.3.1 Lambda Phage Decision Circuit

Lambda phage is a virus that infects the bacterium E. Coli. It consists of a head containing the virus DNA and a tail, which is used to attach to the surface of E. Coli. Such viruses are called bacteriophages and have no own metabolism, but use that of their host instead. On infection, the virus particle binds to the membrane of the bacterium and injects its DNA. Afterwards the host’s metabolism is changed by the gene expression products of the virus. The infected cell may then enter the lytic cycle, which means that new phages are synthesized in the cell, which finally bursts (lysis) to release the new phages. It may also be that the host cell does not enter the lytic cycle, but integrates the virus genome into its own genome. In this case

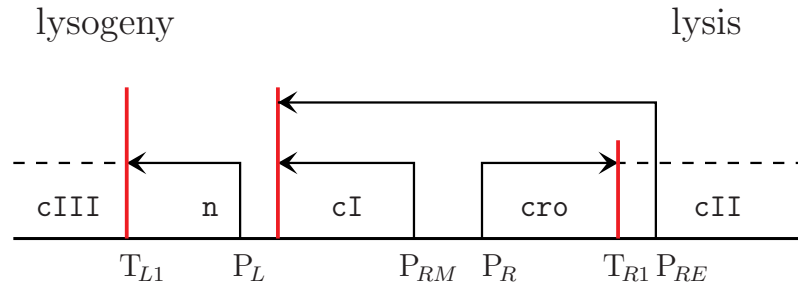


Figure 1.5: Arrangement of parts of the virus genome.

the cell enters the lysogenic cycle. Through cell division, the virus DNA is replicated and in this way the DNA of the daughter cells also contains the integrated virus DNA. Any cell in the lysogenic cycle may, under certain conditions, eventually enter the lytic cycle.

The decision between lysis and lysogeny after infection depends on the temporal pattern of two proteins, **CI** and **Cro**². Their promoters P_{RM} and P_R are arranged in the phage lambda genome as illustrated in Fig. 1.5. On infection, only the promoters P_R and P_L are active. All others have a low basal activation. This leads to an accumulation of **N** and **Cro** proteins (see termination points for RNA Polymerase in Fig. 1.5 that are marked in red color). The protein **N** is used to antiterminate RNA Polymerase upstream from the termination points T_{L1} and T_{R1}. Then transcription of the genes cII, cIII, etc is activated.

Lysis. It is most likely that the cell enters the lytic cycle if the following steps occur next:

- The concentration of **Cro** raises quickly and a negative feedback loop is entered, that is, **Cro** dimers repress P_R and the concentration of **Cro** becomes stable (see also the detailed model of the lysis-lysogeny decision circuit).
- **Cro** dimers also repress P_{RM} and P_L, which leads to a low concentration of **CI**.
- P_{RE} is not activated.
- The transcription of the virus DNA is mostly performed in the rightward direction.

²We use names starting with upper case letters to denote proteins and we use the same names for the corresponding genes, but start with lower case letters.

Lysogeny. It is most likely that the cell enters the lysogenic cycle if the following steps occur next after protein N has antiterminated T_{L1} and T_{R1} :

- The concentration of **Cro** raises slower than that of **CI** (roughly $\#CI$ dimers >145 and $\#Cro$ dimers <55).
- The number of **CI** dimers stabilizes at 140-200 molecules and the dimers repress P_R and P_L , but activate P_{RM} .
- The number of **Cro** and **N** molecules decreases, which makes leftward transcription more likely.

The concentration of **CI** can raise quickly at the beginning since P_{RE} is activated by **CII** molecules (even though their concentration may be low). Later, **CI** molecules are produced via the positive autoregulation at P_{RM} , that is, P_{RM} is activated by **CI** dimers.

The decision between lysis and lysogeny is biased by the following factors:

- (i) the nutritional state of the host cell,
- (ii) the multiplicity of infection (number of phages that are simultaneously infecting the same cell),
- (iii) the cell volume.

At a high level of nutrition the overall proteolytic activity is high, which results in a shorter lifetime of **CII** and **CIII**. This decreases the probability of P_{RE} activation and thus the probability of lysogeny. A high multiplicity of infection increases the number of genes of **CII** and **CIII**. Then P_{RE} is activated at an early stage to kickstart the production of **CI**, which results in a higher probability of lysogeny. For (iii), it has been observed that smaller cells go more likely into lysogeny. Although the exact mechanisms are unknown, the conjecture is that since in small cells the concentrations of **CII** and **CIII** is higher than in large cells, P_{RE} activation is more likely in small cells.

1.3.2 Intrinsic and Extrinsic Molecular Noise

In the previous section we have seen that biological processes are significantly influenced by combinations of random microscopic events in the cell. It remains the question what the origin of this “cellular noise” is. In order to distinguish different sources of noise, the following experiment has been carried out: In [ELSS02], Elowitz et al. report about strains of *E. coli* with two genes that are expressed under the same conditions. Using fluorescent markers in two different color (yellow and cyan) they could determine the gene expression strength of each gene. As a result they obtained

cells with different combinations of yellow and cyan. In some cells, only one of the two genes was expressed whereas in other both genes were expressed. The work of Elowitz et al. is an experimental proof for so-called *intrinsic noise*, which arises from random microscopic events in the cell, such as the location of molecules or the order of the chemical reactions. As opposed to that *extrinsic noise* arises from the variability in a population of genetically identical cells, that is, from the different amounts of cellular components. Modeling systems that are subject to molecular noise is challenging since the traditional reaction rate equation approach is inappropriate. Instead of a deterministic model that predicts a single outcome (with certainty), stochastic models are required. But besides the fact that a stochastic model can predict probabilities for the different outcomes of an experiment, it must take into account the discreteness of the events in the cell. More precisely, instead of assuming continuous changes for the concentrations of chemical species, the model has to consider discrete “jumps” (compare Figure 1.1).

1.4 Basic Concepts of Probability

“WAS EIN PUNKT, EIN RECHTER WINKEL, EIN KREIS IST, WEISS ICH SCHON VOR DER ERSTEN GEOMETRIESTUNDE, ICH KANN ES NUR NOCH NICHT PRÄZISIEREN. EBENSO WEISS ICH SCHON, WAS WAHRSCHEINLICHKEIT IST, EHE ICH ES DEFINIERT HABE.“ (Hans Freudenthal)

This section gives a short primer on discrete and continuous random variables.

1.4.1 Probabilities

Let us consider chance experiments with a countable number of possible outcomes $\omega_1, \omega_2, \dots$. The set $\Omega = \{\omega_1, \omega_2, \dots\}$ of all outcomes is called the *sample space*. Subsets of Ω are called *events* and by 2^Ω we denote the set of all events.

Example 4: ROLLING A DIE AND TOSSING A COIN

If we roll a die, the set of possible outcomes is $\Omega = \{1, 2, \dots, 6\}$. The event “number is even” is given by $E = \{2, 4, 6\}$.

If we toss a coin and count the number of trials until a head turns up for the first time, $\Omega = \{1, 2, \dots\}$. The event “number of trials greater than 10” is given by $E = \{11, 12, \dots\}$.

We define what a *probability* is using Kolmogorov’s axioms.

Definition 1: PROBABILITY

Assume Ω is a discrete (i.e. finite or countably infinite) and non-empty sample space. Let P be a function such that $P : 2^\Omega \rightarrow [0, 1]$. The value $P(A)$ is called the probability of event A if P is such that

1. $P(\Omega) = 1$,
2. for pairwise disjoint events A_1, A_2, \dots it holds that

$$P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots$$

When we reason about the probability of certain events, we can use many arguments from set theory. For instance, if the events A, B , and C are as illustrated in Fig. 1.6, it holds that

$$P(A \cup B) = P(A) + P(B) \text{ and}$$

$$P(A \cup C) = P(A) + P(C) - P(A \cap C).$$

Example 5: ROLLING A DIE

The probability of the events $\{2, 4, 6\}$, $\{2\}$, and $\{1, 2, \dots, 6\}$ are

$$P(\{2, 4, 6\}) = 1/2,$$

$$P(\{2\}) = 1/6, \text{ and}$$

$$P(\{1, 2, \dots, 6\}) = 1, \text{ respectively.}$$

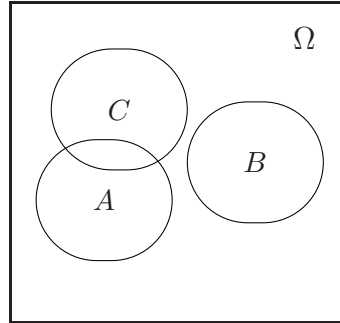


Figure 1.6: Using set arguments for the calculation of event probabilities.

TOSSING A COIN UNTIL HEADS FOR THE FIRST TIME

The probability of the events “exactly n trials” and “more than three trials” are

$$P(\{n\}) = (1/2)^n \text{ and}$$

$$P(\{4, 5, \dots\}) = P(\Omega \setminus \{1, 2, 3\}) = 1 - (1/2 + 1/4 + 1/8) = 1/8.$$

Note that $P(\Omega) = \sum_{\omega \in \Omega} P(\{\omega\}) = 1/2 + 1/4 + 1/8 + \dots = 1$.

The triple $(\Omega, 2^\Omega, P)$ is called a *discrete probability space*.

Definition 2: CONDITIONAL PROBABILITY

Let A, B be events and $P(B) > 0$. Then

$$P(A|B) := \frac{P(A \cap B)}{P(B)}$$

is called probability of A under the condition B . Clearly, this implies that $P(A \cap B) = P(A|B) \cdot P(B)$.

It is easy to show that the value $P(A|B)$ is a probability in the sense of Definition 1.

Example 6: LUNG CANCER

We define the events

A : person gets lung cancer with $P(A) = \frac{72}{200000} = 0.00036$,

B : person is a smoker with $P(B) = 0.25$.

From the people that get lung cancer, 90% are smokers. The experiment consists in choosing a person at random. For the probability of getting lung cancer under the condition of being a smoker, we calculate

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.9 \cdot 0.00036}{0.25} = 0.001296.$$

For the probability of getting lung cancer under the condition of not being a smoker, we get

$$P(A|\bar{B}) = \frac{P(A \cap \bar{B})}{P(\bar{B})} = \frac{0.1 \cdot 0.00036}{0.75} = 0.000048.$$

Thus, the chance of getting lung cancer is around 30 times higher for smokers compared to non-smokers.

Definition 3: INDEPENDENCE

Let $0 < P(B) < 1$. The event A is called independent of B if

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

Example 7: INDEPENDENT EVENTS

Consider the case $A = \Omega$. The event Ω is independent of any event B with $0 < P(B) < 1$ since

$$P(\Omega|B) = \frac{P(\Omega \cap B)}{P(B)} = 1 = \frac{P(\Omega \cap \bar{B})}{P(\bar{B})} = P(\Omega|\bar{B}).$$

Assume now that $\Omega = \{1, 2, \dots, 6\}$, $A = \{5, 6\}$, and $B = \{2, 4, 6\}$. Then

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{6\})}{P(\{2, 4, 6\})} = 2 \cdot \frac{1}{6} = 1/3$$

and

$$P(A|\bar{B}) = \frac{P(A \cap \bar{B})}{P(\bar{B})} = \frac{P(\{5\})}{P(\{1, 3, 5\})} = 2 \cdot \frac{1}{6} = 1/3,$$

which means that $A = \{5, 6\}$ is independent of $B = \{2, 4, 6\}$.

Assume that we have a finite or countably infinite number of events A_1, A_2, \dots that are pairwise disjoint. Assume further $\Omega = A_1 \cup A_2 \cup \dots$ and $P(A_i) > 0$ for all i . Often, the probability of some event B is unknown, but the conditional probabilities $P(B|A_i), \dots$ are known. In this case, $P(B)$ can be computed using the *law of total probability*, which states that

$$P(B) = \sum_i \underbrace{P(B|A_i) \cdot P(A_i)}_{=P(B \cap A_i)}.$$

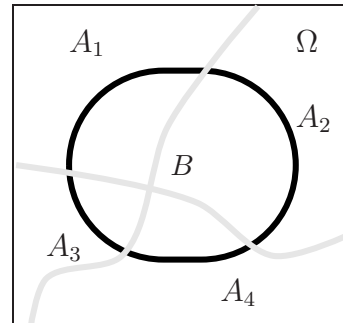


Figure 1.7: The law of total probability.

The corresponding partitioning of Ω is illustrated in Fig. 1.7.

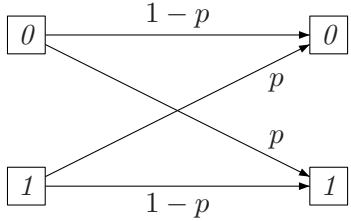
If, however, $P(A_i|B)$ is the probability of interest, one can use *Bayes' theorem*, which states that

$$P(A_i|B) = \frac{P(B \cap A_i)}{P(B)} = \frac{P(B|A_i) \cdot P(A_i)}{P(B)} = \frac{P(B|A_i) \cdot P(A_i)}{\sum_i P(B|A_i) \cdot P(A_i)},$$

where A_1, A_2, \dots are as above.

Example 8: NOISY CHANNEL

Consider the noisy binary channel illustrated below.



If the channel is noise-free ($p = 0$), zero is transmitted from the upper left node to the upper right node. Similarly, one is transmitted from the lower left node to the lower right node. If the channel is noisy, with probability $p > 0$ one is transmitted instead of zero and zero instead of one. Assume that the probability of sending zero is π_0 and the probability of sending one is $\pi_1 = 1 - \pi_0$.

We define the events

- A_0 : send zero with $P(A_0) = \pi_0$,
- A_1 : send one with $P(A_1) = \pi_1 := 1 - \pi_0$,
- B_0 : receive zero,
- B_1 : receive one,

Then, $P(B_1|A_0) = P(B_0|A_1) = p$ and $P(B_1|A_1) = P(B_0|A_0) = 1 - p$. We calculate

$$\begin{aligned} P(B_1) &= P(B_1|A_0) \cdot P(A_0) + P(B_1|A_1) \cdot P(A_1) = p \cdot \pi_0 + (1 - p) \cdot \pi_1, \\ P(B_0) &= P(B_0|A_0) \cdot P(A_0) + P(B_0|A_1) \cdot P(A_1) = (1 - p) \cdot \pi_0 + p \cdot \pi_1. \end{aligned}$$

1.4.2 Discrete Random Variables

A random variable is used to represent an outcome of an experiment. Technically, the fact that this variable is “random” (that is, it takes values with a certain probability), is realized by using a mapping.

Definition 4: DISCRETE RANDOM VARIABLE

Let $(\Omega, 2^\Omega, P)$ be a discrete probability space. A function

$$X : \Omega \rightarrow \mathbb{R}$$

is called a discrete (real-valued) random variable on $(\Omega, 2^\Omega, P)$.

Now, that we mapped the outcome of an experiment to \mathbb{R} , we need to assign probabilities to the subsets of $X(\Omega) = \{a \in \mathbb{R} \mid \exists \omega \in \Omega : X(\omega) = a\}$. Since we already have probabilities for events $A \subseteq \Omega$, we define a function $P_X : 2^{X(\Omega)} \rightarrow [0, 1]$ such that, for $A \in 2^{X(\Omega)}$,

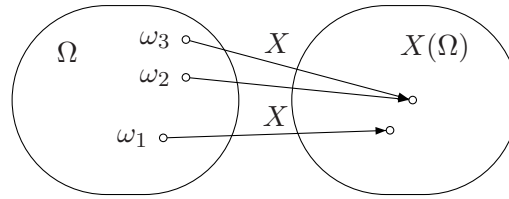


Figure 1.8: Relating random variable and probability.

$$P_X(A) := P(X^{-1}(A)) = P(\{\omega \in \Omega \mid X(\omega) \in A\}).$$

Then $(X(\Omega), 2^{X(\Omega)}, P_X)$ is a discrete probability space.

Example 9: ROLLING TWO DICE

Assume $\Omega = \{(\omega_1, \omega_2) \mid \omega_1, \omega_2, \in \{1, \dots, 6\}\}$ and $X : \Omega \rightarrow \mathbb{R}$ is such that $X(\omega_1, \omega_2) = \omega_1 + \omega_2$.

Then, the probability of having the number 10 is

$$\begin{aligned} P_X(\{10\}) &= P(X^{-1}(\{10\})) = P(\{(\omega_1, \omega_2) \in \Omega \mid X(\omega_1, \omega_2) = 10\}) \\ &= P(\{(4, 6)\}) + P(\{(6, 4)\}) + P(\{(5, 5)\}) = 1/12. \end{aligned}$$

In the sequel, we use the following “shortcuts“ to refer to subset of Ω :

- “ $X = a$ ” stands for the set $\{\omega \in \Omega \mid X(\omega) = a\}$
- “ $X \leq a$ ” stands for the set $\{\omega \in \Omega \mid X(\omega) \leq a\}$
- “ $X < a$ ” ...

Thus, for instance,

$$P(X \leq a) = P(\{\omega \in \Omega \mid X(\omega) \leq a\}) = \sum_{c \in X(\Omega), c \leq a} P(X = c).$$

The function $f : X(\Omega) \rightarrow [0, 1]$ with $f(a) = P(X = a)$ is called the discrete probability distribution of X .

Definition 5: CUMULATIVE PROBABILITY DISTRIBUTION

Let X be a discrete (real-valued) random variable. The function $F : \mathbb{R} \rightarrow [0, 1]$ with

$$F(x) := P(X \leq x) = \sum_{a \in X(\Omega), a \leq x} P(X = a).$$

is called the cumulative probability distribution of X .

Example 10: ROLLING TWO DICE

Assume that X is defined as in Example 9. The discrete probability distribution and the cumulative probability distribution of X are shown in Figure 1.10.

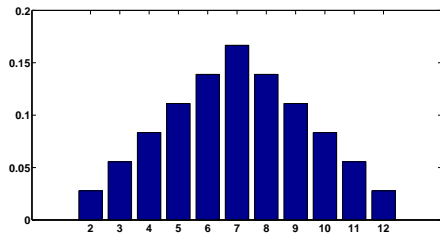


Figure 1.9: Discrete probability distribution (rolling two dice).

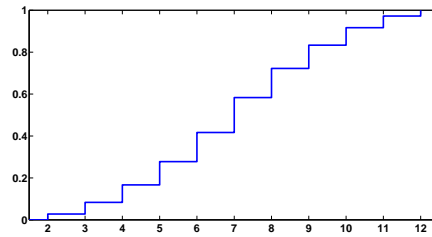


Figure 1.10: Cumulative probability (rolling two dice).

Besides the probability distribution of a discrete random variable X , other values related to X are of interest. The expectation and variance of X are defined as

- $E(X) = \sum_{x \in X(\Omega)} x \cdot P(X = x)$
- $V(X) = \sum_{x \in X(\Omega)} (x - E(X))^2 \cdot P(X = x)$

respectively. (Note that the sum might not converge, in which case the expectation/variance does not exist.) The standard deviation of X is given by $\sqrt{V(X)}$.

Example 11: EXPECTED VALUE OF SPECIAL RANDOM VARIABLES

Let $c \in \mathbb{R}$. Assume $X(\omega) = c$ for all $\omega \in \Omega$. Then $E(X) = c$.

Let $A \subseteq \Omega$ and define $Y = I_A$ where

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{otherwise.} \end{cases}$$

Then $E(Y) = 1 \cdot P(A) + 0 \cdot P(\bar{A}) = P(A)$.

It is possible to connect random variables on the same sample space Ω using operations such as $+$, $-$, \cdot , etc. For instance, we define $Z = X + Y$ as a random variable on Ω by

$$Z(\omega) = X(\omega) + Y(\omega) \text{ for all } \omega \in \Omega.$$

The probability $P(Z = z) = P(X + Y = z)$ is then well-defined since

$$P(Z = z) = P(\omega \in \Omega \mid Z(\omega) = z) = \sum_{\omega \in \Omega, X(\omega) + Y(\omega) = z} P(\{\omega\}).$$

Moreover,

$$\sum_{z \in Z(\Omega)} P(Z = z) = \sum_{z \in Z(\Omega)} \sum_{\omega \in \Omega, X(\omega) + Y(\omega) = z} P(\{\omega\}) = P(\Omega) = 1.$$

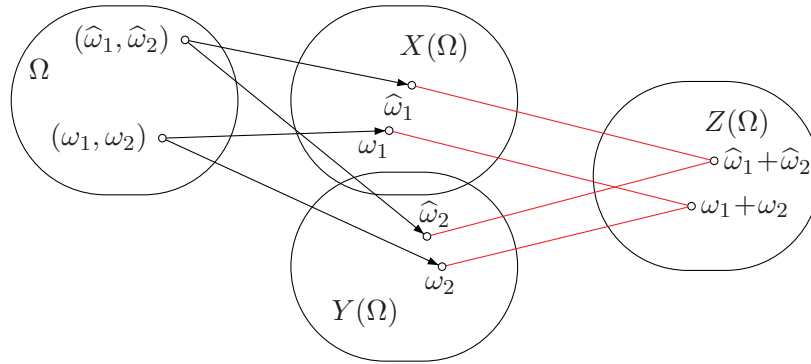


Figure 1.11: Sum of two random variables (rolling two dice).

Example 12: ROLLING TWO DICE

Assume that $\Omega = \{1, 2, \dots, 6\}^2$ and $X, Y : \Omega \rightarrow \mathbb{R}$ are such that, for $(\omega_1, \omega_2) \in \Omega$,

$$X(\omega_1, \omega_2) = \omega_1, \quad Y(\omega_1, \omega_2) = \omega_2.$$

Then for $Z = X + Y$ we have

$$Z(\omega_1, \omega_2) = X(\omega_1, \omega_2) + Y(\omega_1, \omega_2) = \omega_1 + \omega_2$$

and (see illustration in Fig. 1.11)

$$P(Z = z) = \sum_{\substack{(\omega_1, \omega_2) \in \Omega, \\ X(\omega_1, \omega_2) + Y(\omega_1, \omega_2) = z}} P(\{(\omega_1, \omega_2)\}) = \sum_{\substack{(\omega_1, \omega_2) \in \Omega, \\ \omega_1 + \omega_2 = z}} P(\{(\omega_1, \omega_2)\}).$$

Definition 6: JOINT PROBABILITY DISTRIBUTION

Let X, Y be discrete (real-valued) random variables on the same probability space $(\Omega, 2^\Omega, P)$. The function $P_{X,Y} : X(\Omega) \times Y(\Omega) \rightarrow [0, 1]$ with

$$P_{X,Y}(a, b) := P(X = a \wedge Y = b) = P(X^{-1}(\{a\}) \cap Y^{-1}(\{b\})).$$

is called the joint probability distribution of X and Y .

Definition 7: INDEPENDENCE OF RANDOM VARIABLES

Let X, Y be discrete (real-valued) random variables on the same probability space $(\Omega, 2^\Omega, P)$. We call X and Y independent iff for all $a \in X(\Omega)$, $b \in Y(\Omega)$

$$P(X = a \mid Y = b) = P(X = a).$$

(Or equivalently, X and Y are independent iff $P(X = a \wedge Y = b) = P(X = a) \cdot P(Y = b)$.)

Example 13: ROLLING TWO DICE

Assume that X and Y are defined as in Example 12. Then X and Y are independent, since for any $a, b \in \{1, \dots, 6\}$,

$$\begin{aligned} P(X = a \wedge Y = b) &= P(X^{-1}(\{a\}) \cap Y^{-1}(\{b\})) \\ &= P(\{(\omega_1, \omega_2) \in \Omega \mid \omega_1 = a\} \cap \{(\omega_1, \omega_2) \in \Omega \mid \omega_2 = b\}) \\ &= P(\{(a, b)\}) = 1/36 \end{aligned}$$

and

$$\begin{aligned} P(X = a) \cdot P(Y = b) &= P(X^{-1}(\{a\})) \cdot P(Y^{-1}(\{b\})) \\ &= P(\{(\omega_1, \omega_2) \in \Omega \mid \omega_1 = a\}) \cdot P(\{(\omega_1, \omega_2) \in \Omega \mid \omega_2 = b\}) \\ &= 6/36 \cdot 6/36 = 1/36. \end{aligned}$$

We are now able to state some properties of the expectation. Let X, Y be discrete random variables on the same probability space and assume that $E(X)$ and $E(Y)$ exist. Then, for $a, b \in \mathbb{R}$

1. $E(X) = \sum_{\omega \in \Omega} X(\omega)P(\{\omega\})$,
2. $E(a \cdot X + b) = a \cdot E(X) + b$,
3. $E(X + Y) = E(X) + E(Y)$.
4. If X and Y are independent, then $E(X \cdot Y) = E(X) \cdot E(Y)$.

Proof. (1.)

$$\begin{aligned} E(X) &= \sum_{x \in X(\Omega)} x \cdot P(X = x) && \text{(definition of } E(X)) \\ &= \sum_{x \in X(\Omega)} x \cdot P(\{\omega \mid X(\omega) = x\}) && \text{(definition of } P(X = x)) \\ &= \sum_{x \in X(\Omega)} x \cdot \sum_{\omega \in \Omega \wedge X(\omega) = x} P(\{\omega\}) && \text{(2nd axiom of Def. 1)} \\ &= \sum_{\omega \in \Omega} X(\omega)P(\{\omega\}) && \text{(} X \text{ is a function)} \end{aligned}$$

□

Proof. (2.) Let $Z = a \cdot X + b$ be a discrete random variable on the same

probability space as X , such that $Z(\omega) = a \cdot X(\omega) + b$. We now find:

$$\begin{aligned}
 E(a \cdot X + b) &= E(Z) && \text{(definition of } Z) \\
 &= \sum_{\omega \in \Omega} Z(\omega)P(\{\omega\}) && \text{(first property)} \\
 &= \sum_{\omega \in \Omega} (aX(\omega) + b) \cdot P(\{\omega\}) && \text{(definition of } Z) \\
 &= a \sum_{\omega \in \Omega} X(\omega) \cdot P(\{\omega\}) + b \sum_{\omega \in \Omega} P(\{\omega\}) && \text{(calculus)} \\
 &= a \sum_{\omega \in \Omega} X(\omega) \cdot P(\{\omega\}) + bP(\Omega) && \text{(2nd axiom of Def. 1)} \\
 &= a \sum_{\omega \in \Omega} X(\omega) \cdot P(\{\omega\}) + b && \text{(1st axiom of Def. 1)} \\
 &= a \cdot E(X) + b && \text{(first property)}
 \end{aligned}$$

□

Proof. (3.) Let $Z = X + Y$ be a discrete random variable on the same probability space as X and Y , such that $Z(\omega) = X(\omega) + Y(\omega)$. We now find:

$$\begin{aligned}
 E(X + Y) &= E(Z) && \text{(definition of } Z) \\
 &= \sum_{\omega \in \Omega} Z(\omega)P(\{\omega\}) && \text{(first property)} \\
 &= \sum_{\omega \in \Omega} (X(\omega) + Y(\omega)) \cdot P(\{\omega\}) && \text{(definition of } Z) \\
 &= \sum_{\omega \in \Omega} X(\omega) \cdot P(\{\omega\}) + \sum_{\omega \in \Omega} Y(\omega) \cdot P(\{\omega\}) && \text{(calculus)} \\
 &= E(X) + E(Y) && \text{(first property)}
 \end{aligned}$$

□

Proof. (4.) Let $Z = X \cdot Y$ be a discrete random variable on the same

probability space as X and Y , such that $Z(\omega) = X(\omega) \cdot Y(\omega)$. We now find:

$$\begin{aligned}
 & E(X \cdot Y) \\
 &= E(Z) && \text{(definition of } Z\text{)} \\
 &= \sum_{\omega \in \Omega} Z(\omega) P(\{\omega\}) && \text{(first property)} \\
 &= \sum_{\omega \in \Omega} X(\omega) \cdot Y(\omega) P(\{\omega\}) && \text{(definition of } Z\text{)} \\
 &= \sum_{x \in X(\Omega) \wedge y \in Y(\Omega)} x \cdot y \cdot \sum_{\omega \in \Omega \wedge X(\omega)=x \wedge Y(\omega)=y} P(\{\omega\}) \\
 & && \text{(} X, Y \text{ are functions)} \\
 &= \sum_{x \in X(\Omega) \wedge y \in Y(\Omega)} x \cdot y \cdot P(\{\omega \mid X(\omega) = x \wedge Y(\omega) = y\}) \\
 & && \text{(2nd axiom of Def. 1)} \\
 &= \sum_{x \in X(\Omega) \wedge y \in Y(\Omega)} x \cdot y \cdot P(X = x \wedge Y = y) \\
 & && \text{(definition of } P(\dots)\text{)} \\
 &= \sum_{x \in X(\Omega) \wedge y \in Y(\Omega)} x \cdot y \cdot P(X = x) \cdot P(Y = y) \\
 & && \text{(independence of } X \text{ and } Y\text{)} \\
 &= \sum_{x \in X(\Omega)} x \cdot P(X = x) \cdot \sum_{y \in Y(\Omega)} y \cdot P(Y = y) \\
 & && \text{(calculus)} \\
 &= E(X) \cdot E(Y) && \text{(definition of } E(X), E(Y)\text{)}
 \end{aligned}$$

□

We list the properties of the variance operator without proof: Let X, Y be discrete random variables on the same probability space and assume that $E(X)$, $E(Y)$, $VAR(X)$, and $VAR(Y)$ exist. Then

1. $VAR(a \cdot X + b) = a^2 \cdot VAR(X)$ for $a, b \in \mathbb{R}$,
2. $VAR(X + Y) = VAR(X) + VAR(Y) + 2 \cdot (E(X \cdot Y) - E(X) \cdot E(Y))$.
3. If X and Y are independent, then $VAR(X + Y) = VAR(X) + VAR(Y)$.

The definition of independent random variables can be extended to more than two variables as follows.

Definition 8: INDEPENDENCE (OF n DISCRETE RANDOM VARIABLES)

For $n \in \mathbb{N}$, let X_1, \dots, X_n be discrete (real-valued) random variables on the same probability space $(\Omega, 2^\Omega, P)$. We call X_1, \dots, X_n independent iff for all $a_1 \in X_1(\Omega), \dots, a_n \in X_n(\Omega)$

$$P(X_1 = a_1, \dots, X_n = a_n) = P(X_1 = a_1) \cdot \dots \cdot P(X_n = a_n).$$

Note that if X_1, \dots, X_n are independent, then

$$E(X_1 \cdot \dots \cdot X_n) = E(X_1) \cdot \dots \cdot E(X_n)$$

and

$$\text{VAR}(X_1 + \dots + X_n) = \text{VAR}(X_1) + \dots + \text{VAR}(X_n).$$

Let us now consider some important distributions.

Example 14: GEOMETRIC DISTRIBUTION

Consider an experiment where the probability of an event A is $P(A) = p$. We repeat this experiment until A occurs for the first time. Let X be the random variable that describes the number of trials until A occurs for the first time. Then X is called geometrically distributed. We have $P(X = i) = (1-p)^{i-1} \cdot p$ and $E(X) = 1/p$. The variance of X is $V(X) = (1-p)/p$.

Example 15: POISSON DISTRIBUTION

Consider a call center where on average $\mu = 6$ calls per minute arrive. Let X be the random variable that represents the number of calls in the next minute and assume $P(X = k) = \frac{\mu^k}{k!} e^{-\mu}$. Then X is called Poisson distributed and has expectation

$$E(X) = \sum_{k=0}^{\infty} k \cdot \frac{\mu^k}{k!} e^{-\mu} = \mu e^{-\mu} \sum_{k=1}^{\infty} \frac{\mu^{k-1}}{(k-1)!} = \mu e^{-\mu} e^{\mu} = \mu$$

and variance $V(X) = \mu$ (without proof).

1.4.3 Continuous Random Variables

Consider a chance experiment where the sample space Ω contains uncountably many elements. In this case, assigning probabilities to all elements in 2^Ω poses problems. Assume, for instance, we randomly choose a real number in $[0, 1]$. If all numbers are equally likely to occur, we have to assign probability zero to each since their “sum” must be one (note that the sum over uncountably many nonzero values is undefined). Instead of giving a new definition of probabilities, we restrict ourselves to a set of events, called σ -algebra, for which we can define probabilities as in the discrete case.

If we define a random variable as a function $X : \Omega \rightarrow \mathbb{R}$, we want to reason about the probability of events such as $X = x$ for any $x \in \mathbb{R}$ or $a < X \leq b$ for any interval $(a, b]$. From the properties of probabilities (see Def. 1), we then know the probability of the disjoint union of countably many of such sets as well as the probability of the complement of such a set³.

Definition 9: σ -ALGEBRA

A set $\mathcal{F} \subseteq 2^\Omega$ is called a σ -algebra if

1. $\Omega \in \mathcal{F}$,
2. $A \in \mathcal{F}$ implies $\bar{A} \in \mathcal{F}$.

³From the two conditions in Def. 1, one can easily derive that $P(\bar{A}) = 1 - P(A)$.

3. If $A_1, A_2, \dots \in \mathcal{F}$ is a sequence of sets then

$$A_1 \cup A_2 \cup \dots \in \mathcal{F}.$$

The most important example of a σ -algebra, which is needed in the sequel, is the σ -algebra that is generated by a set $\mathcal{E} \subseteq 2^\Omega$. We define the smallest σ -algebra that contains \mathcal{E} by

$$\sigma(\mathcal{E}) := \cap \{ \mathcal{F} \supset \mathcal{E} : \mathcal{F} \text{ is a } \sigma\text{-algebra} \}.$$

Example 16: GENERATED σ -ALGEBRA

For simplicity, we consider the finite set $\Omega = \{1, 2, \dots, 6\}$. Let $\mathcal{E} = \{\{2, 6\}, \{5, 6\}\}$. Then

$$\sigma(\mathcal{E}) = \{\{2, 6\}, \{5, 6\}, \{1, 3, 4, 5\}, \{1, 2, 3, 4\}, \{1, 2, 3, 4, 5\}, \{6\}, \dots\}.$$

If Ω is finite, the idea is to construct a sequence of sets in an iterative fashion, where we start with \mathcal{E} and obtain the next set from the previous one by joining and complementing elements of the current set. If no new element can be constructed by union or complement operations, the current set equals $\sigma(\mathcal{E})$.

Example 17: BOREL σ -ALGEBRA

Assume $\Omega = \mathbb{R}$ and $\mathcal{E} = \{(a, b] : a, b \in \mathbb{R}, a \leq b\}$. The σ -algebra $\mathcal{B} := \sigma(\mathcal{E})$ is called the Borel algebra on the reals. It contains all subsets, called Borel sets, of $2^{\mathbb{R}}$ that can be obtained from \mathcal{E} by countable union and complement operations. Note that also intervals of the form (a, b) or $[a, b]$ are Borel sets. A similar construction is possible for $\Omega = \mathbb{R}^n$. Intuitively, the Borel sets are those sets, for which we can assign a “volume”, “area” or “size”. Subsets of \mathbb{R}^n that are not Borel sets are only of theoretical interest since for practical applications they are not of importance.

We are now able to give a more general definition of a probability space (i.e. also for sample sets with uncountably many elements).

Definition 10: PROBABILITY SPACE

Let Ω be a nonempty set and let $\mathcal{F} \subseteq 2^\Omega$ be a σ -algebra. A probability space is a triple (Ω, \mathcal{F}, P) where the probability measure $P : \mathcal{F} \rightarrow [0, 1]$ is such that

- $P(\Omega) = 1$ and,
- if $A_1, A_2, \dots \in \mathcal{F}$ is a sequence of pairwise disjoint sets, then

$$P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots$$

Example 18: DISCRETE PROBABILITY SPACE

Any discrete probability space $(\Omega, 2^\Omega, P)$ fulfils Def. 10, since 2^Ω is a σ -algebra and P is as in Def. 1.

Example 19: ONE-DIMENSIONAL INTERVAL

Let $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}$ and $0 \leq a < b$, $a, b \in \mathbb{R}$. Consider a probability measure $P : \mathcal{F} \rightarrow [0, 1]$ such that

$$P(\{\omega \in \Omega \mid x < \omega \leq y\}) = \begin{cases} \frac{\min(y,b) - \max(x,a)}{b-a} & \text{if } \max(x, a) < \min(y, b) \\ 0 & \text{otherwise.} \end{cases}$$

Similar to how we extended the set $\mathcal{E} = \{(a, b] : a, b \in \mathbb{R}, a \leq b\}$ to a σ -algebra, we can show that if P is a probability measure and defined as above for all half-open intervals, its value for the remaining sets in \mathcal{B} is uniquely determined.

Definition 11: REAL-VALUED RANDOM VARIABLE

Let (Ω, \mathcal{F}, P) be a probability space. A real-valued random variable on (Ω, \mathcal{F}, P) is a function $X : \Omega \rightarrow \mathbb{R}$ such that for all $A \in \mathcal{B}$

$$X^{-1}(A) = \{\omega \in \Omega \mid X(\omega) \in A\} \in \mathcal{F}.$$

The above definition ensures that if we want to know the probability that X is in some Borel set A , we can consider the inverse image of A with respect to X , for which we know its probability.

Clearly, we can define a probability measure $P_X : \mathcal{B} \rightarrow [0, 1]$ by setting $P_X(A) := P(\{\omega \mid X(\omega) \in A\}) = P(X^{-1}(A))$ and use similar notations as in the discrete case (e.g. $P(a < X \leq b)$).

Definition 12: CUMULATIVE PROBABILITY DISTRIBUTION

Let X be a real-valued random variable on (Ω, \mathcal{F}, P) . The function $F : \mathbb{R} \rightarrow [0, 1]$ with $x \mapsto F(x) := P(X \leq x)$ is called the cumulative probability distribution of X .

Example 20: ONE-DIMENSIONAL INTERVAL

Assume that X is a randomly chosen point in the interval $[a, b]$ and (Ω, \mathcal{F}, P) is as in Ex. 19. Then

$$F(y) = P(X \leq y) = \begin{cases} \frac{y-a}{b-a} & \text{if } y \in [a, b], \\ 1 & \text{if } y > b, \\ 0 & \text{otherwise.} \end{cases}$$

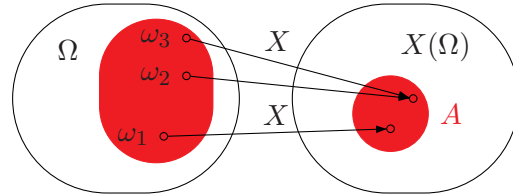


Figure 1.12: Is the inverse image of A w.r.t. X an element of \mathcal{F} ?

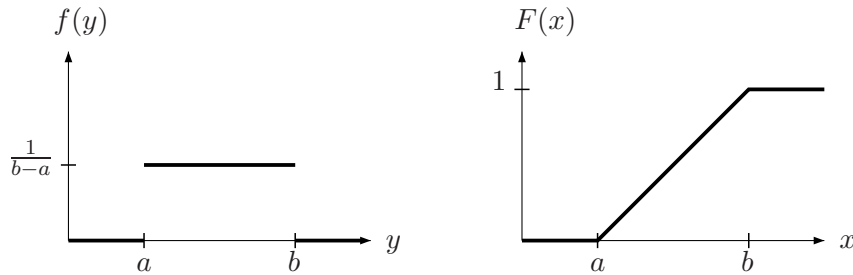


Figure 1.13: Density and cumulative probability distribution of a random variable X .

We call a random variable X on (Ω, \mathcal{F}, P) *discrete* if $X(\Omega)$ is a discrete set (finite or countably infinite). We call X *continuous* if $X(\Omega)$ contains uncountably many elements and there exists a non-negative and integrable function $f : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$, called *density*, with

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(y) dy.$$

Clearly,

$$\int_{-\infty}^{\infty} f(y) dy = 1$$

since $F(\infty) = 1$, but note that $P(X = y) \neq f(y)$.

Example 21: ONE-DIMENSIONAL INTERVAL

Assume that X is as in Ex. 20. Then X is a continuous random variable since the (constant) function f with

$$f(y) = \begin{cases} \frac{1}{b-a} & \text{if } y \in [a, b], \\ 0 & \text{otherwise.} \end{cases}$$

is the density of X . We can verify this by calculating

$$F(x) = \int_{-\infty}^x f(y) dy = \int_a^x \frac{1}{b-a} dy = \left[\frac{y}{b-a} \right]_a^x = \frac{x-a}{b-a} = P(X \leq x)$$

for $x \in [a, b]$, $F(x) = 0$ for $x < a$, and $F(x) = 1$ for $x > b$. Figure 1.13 shows a plot of the functions f and F .

The expectation and variance of a continuous random variable X with density f are defined as

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx \text{ and } V(X) = \int_{-\infty}^{\infty} (x - E(X))^2 \cdot f(x) dx.$$

Note that these integrals may not exist, in which case the expectation/variance is undefined.

Example 22: EXPONENTIAL DISTRIBUTION

Let $\lambda > 0$. We say that a continuous random variable X is exponentially distributed with parameter λ if the density of X is such that, for $t \in \mathbb{R}$,

$$f(t) = \begin{cases} \lambda \cdot e^{-\lambda t} & \text{if } t \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

The cumulative probability distribution of X is then given by

$$F(x) = \begin{cases} \int_{-\infty}^x f(t) dt = \int_0^x \lambda \cdot e^{-\lambda t} dt = 1 - e^{-\lambda x} & \text{if } t \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Many results for discrete random variables carry over to the continuous setting. For instance, properties of the expectation and variance (e.g. $E(X + Y) = E(X) + E(Y)$) or results concerning the connection of random variables, joint distributions, stochastic independence, etc. We do not discuss them here, as they are very similar to the results presented above. Moreover, in the sequel, we will mostly work with discrete random variables.

1.5 Stochastic Chemical Kinetics

Consider a reaction volume with chemical species S_1, \dots, S_N and chemical reaction types R_1, \dots, R_M . We assume that each reaction is elementary and that we never have more than two reactants (bimolecular). Note that both assumptions are harmless since we can describe any other reaction as the composition of elementary bimolecular reactions. Moreover, collisions between three or more molecules are extremely rare.

An ideal model of the time evolution of the system would track the exact positions and velocities of all molecules. Whenever molecules collide, chemical reactions may occur. Unfortunately, such a model is infeasible for nearly all systems. Moreover, the data that is available about the system is not detailed enough.

As a more abstract model, we assume the positions and velocities are given by probability distributions. For the positions we assume that the reaction volume is well-stirred, which means that the molecules are uniformly distributed over the reaction volume. For the velocities we assume that the reaction volume is in thermal equilibrium. The velocity of a single molecule is then given by the Maxwell-Boltzmann distribution. Since the position and velocity of each molecule is now known we have that the number of molecules of each species is enough information to predict the future behavior of the

system. That is, given a vector $(\#S_1, \dots, \#S_N)$ for the populations at time instant t , we can determine the populations at a later time instant $t + \Delta$. Under these assumptions molecules collide randomly and chemical reactions will occur at random points in time. Thus, we can define a chance experiment whose outcomes are functions $\omega : \mathbb{R}_{\geq 0} \rightarrow \mathbb{Z}_+^N$ such that $\omega(t) = (x_1, \dots, x_N)$ is the population vector at time $t \geq 0$. The state of the system at time t can then be represented by a vector of random variables (*random vector*)

$$\vec{X}^{(t)} = (X_1^{(t)}, \dots, X_N^{(t)}),$$

where, for $1 \leq i \leq N$, $X_i^{(t)}$ represents the number of molecules of species S_i at time t . Note that $\vec{X}^{(t)}$ is a discrete random vector, since the number of possible values is a subset of the countable

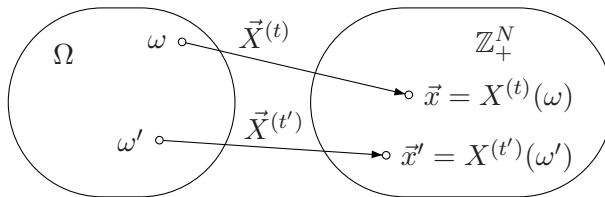
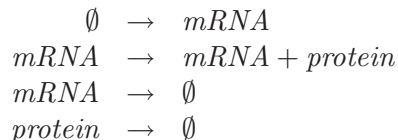


Figure 1.14: Mapping outcomes to states.

set \mathbb{Z}_+^N . On the other hand, we have uncountably many random vectors (since time is continuous) that are all defined on the same probability space (Ω, \mathcal{F}, P) . Such a collection $(\vec{X}^{(t)})_{t \geq 0}$ of random variables (or random vectors) is called a stochastic process and we simply write \vec{X} instead of $(\vec{X}^{(t)})_{t \geq 0}$. In the sequel we will see that we are dealing with a special case of a stochastic process. First of all, the co-domain $\mathbb{Z}_+^N = \{0, 1, \dots\}^N$, called *state space*, is countable. Moreover, for $\omega \in \Omega$, the functions $t \mapsto \vec{X}^{(t)}(\omega)$, called *trajectories*, have a simple jump form as illustrated in Fig. 1.1. Each “jump” corresponds to the occurrence of a chemical reactions. Since $\vec{X}^{(t)}$ represents the state of the system at time t , we have $\vec{X}^{(t)}(\omega) := \omega(t)$, that is, given an outcome ω , the random vector $\vec{X}^{(t)}$ projects ω onto the value of ω at time t (see Fig. 1.14). Note that we can also define Ω in a way that $\omega(t)$ contains more information than only the population numbers. For instance, it could also contain information about the velocities and locations of the molecules. In the sequel, however, we always assume that $\vec{X}^{(t)}(\omega)$ is the population vector at time t (and does not contain further information) and do not specify Ω in detail.

Example 23: GENE EXPRESSION

Consider the following four reactions:



The associated Markov chain has state space \mathbb{Z}_+^2 and an outcome is a func-

tion $\omega : \mathbb{R}_{\geq 0} \rightarrow \mathbb{Z}_+^2$ such that, if $\omega(t) = (\omega_1(t), \omega_2(t))$, on the trajectory $t \mapsto \vec{X}^{(t)}(\omega)$ the number of mRNA molecules at time t is $X_1^{(t)}(\omega) = \omega_1(t)$ and the number of protein molecules at time t is $X_2^{(t)}(\omega) = \omega_2(t)$.

1.5.1 Transition Probabilities

Assume that the process \vec{X} is in state $\vec{x} \in \mathbb{Z}_+^N$ at time t , i.e. we consider the set “ $\vec{X}^{(t)} = \vec{x}$ ” = $\{\omega \mid \omega(t) = \vec{x}\}$ of outcomes where \vec{X} is in state \vec{x} at time t . The process is in state \vec{y} at time $t + \Delta$ with probability

$$P\left(\vec{X}^{(t+\Delta)} = \vec{y} \mid \vec{X}^{(t)} = \vec{x}\right).$$

We call the above values *transition probabilities*. Clearly, we have

$$\begin{aligned} & \sum_{\vec{y} \in \vec{X}^{(t+\Delta)}(\Omega)} P\left(\vec{X}^{(t+\Delta)} = \vec{y} \mid \vec{X}^{(t)} = \vec{x}\right) \\ &= \frac{1}{P(\vec{X}^{(t)} = \vec{x})} \sum_{\vec{y}} P\left(\vec{X}^{(t+\Delta)} = \vec{y}, \vec{X}^{(t)} = \vec{x}\right) \\ &= \frac{1}{P(\vec{X}^{(t)} = \vec{x})} \sum_{\vec{y}} P(\{\omega \mid \omega(t + \Delta) = \vec{y}, \omega(t) = \vec{x}\}) \\ &= \frac{1}{P(\vec{X}^{(t)} = \vec{x})} P(\{\omega \mid \omega(t) = \vec{x}\}) \\ &= \frac{1}{P(\vec{X}^{(t)} = \vec{x})} P(\vec{X}^{(t)} = \vec{x}) = 1. \end{aligned}$$

Since in the abstract model the species’ populations at time t is enough information to predict the future evolution of the system, we have for all $\Delta > 0$, $t \geq 0$, $t_0, t_1, \dots, t_n \in [0, t)$ with $t_0 < \dots < t_n$ and $\vec{x}, \vec{x}_0, \dots, \vec{x}_n \in \mathbb{Z}_+^N$,

$$\begin{aligned} & P\left(\vec{X}^{(t+\Delta)} = \vec{y} \mid \vec{X}^{(t)} = \vec{x}, \vec{X}^{(t_n)} = \vec{x}_n, \dots, \vec{X}^{(t_0)} = \vec{x}_0\right) \\ &= P\left(\vec{X}^{(t+\Delta)} = \vec{y} \mid \vec{X}^{(t)} = \vec{x}\right). \end{aligned}$$

This relationship is called the *Markov property*, named after the Russian mathematician A.A. Markov.

Note that we cannot determine the probability $P(\vec{X}^{(t)} = \vec{x})$ without a base case for the recursion that is given by the transition probabilities, i.e.,

$$\begin{aligned} & P(\vec{X}^{(t)} = \vec{x}) \\ &= \sum_{\vec{x}_n} P\left(\vec{X}^{(t)} = \vec{x} \mid \vec{X}^{(t_n)} = \vec{x}_n\right) \cdot P(\vec{X}^{(t_n)} = \vec{x}_n) \\ &= \sum_{\vec{x}_n, \vec{x}_{n-1}} P\left(\vec{X}^{(t)} = \vec{x} \mid \vec{X}^{(t_n)} = \vec{x}_n\right) \cdot P\left(\vec{X}^{(t_n)} = \vec{x}_n \mid \vec{X}^{(t_{n-1})} = \vec{x}_{n-1}\right) \\ & \quad \cdot P(\vec{X}^{(t_{n-1})} = \vec{x}_{n-1}) \\ &= \dots \end{aligned}$$

In order to uniquely determine the probability measure P , we assume that $P(\vec{X}^{(0)} = \vec{x}_0)$ is given for all $x_0 \in \mathbb{Z}_+^N$, that is, we fix the distribution at time $t = 0$ (and therefore the distributions for all remaining time points are also fixed)⁴.

Change Vectors. Recall that each “jump” of a trajectory $t \mapsto \vec{X}^{(t)}(\omega)$ corresponds to the occurrence of a chemical reaction. For instance, a possible trajectory for the system in Ex. 23 may be the function illustrated in Fig. 1.15 (where we only plot the first component of $\vec{X}^{(t)}(\omega)$, that is, the number of mRNA molecules).

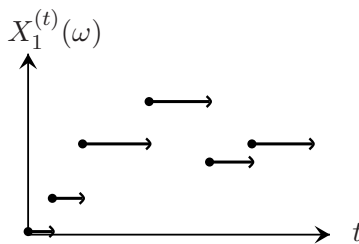


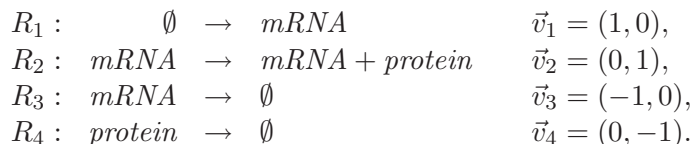
Figure 1.15: State changes occur at discrete points in time.

The effect of the occurrence of a chemical reaction of type R_j , $1 \leq j \leq M$ is given by the *vector of change* $\vec{v}_j \in \{-2, -1, 0, 1, 2\}^N$ of R_j . The i -th entry is the difference between

the number of molecules of species S_i *gained* by R_j and the number of molecules of species S_i *consumed* by R_j .

Example 24: GENE EXPRESSION

The change vectors of the four reactions of the gene expression example are given by



Thus, if \vec{x} is the current state of the system and an instance of reaction R_j occurs, the next state will be $\vec{x} + \vec{v}_j$.

Reactant Combinations. We consider the number of possibilities to combine the reactants of a certain reaction if the current state is $\vec{x} = (x_1, \dots, x_N)$. If the reaction is of the form $S_i \rightarrow products$ there are x_i reactants in the system, which yields x_i different possible instances of the reaction. If the reaction is of the form $S_i + S_k \rightarrow products$, $i \neq k$, there are $x_i \cdot x_k$ different ways to combine the reactants. In the case of $2S_i \rightarrow products$, there are $\binom{x_i}{2} = 0.5 \cdot x_i \cdot (x_i - 1)$ possible combinations of two molecules of type S_i . Finally, for $\emptyset \rightarrow products$, there is only one possible instance of the reaction.

⁴Often we will start with probability one in a single state, i.e. $P(\vec{X}^{(0)} = \vec{x}_0) = 1$ for some state x_0 .

Infinitesimal Transition Probabilities. Assume now that $[t, t + \Delta)$ is an infinitesimal time interval, that is, it “small enough” such that the probability of more than one state change within $[t, t + \Delta)$ is negligible (disappears in the limit). The fundamental premise of stochastic chemical kinetics is that the probability of a type R_j reaction within the next Δ time units is proportional to the product of the number of reactant combinations and Δ , that is, there exists a constant $c_j > 0$ such that

$$P\left(\vec{X}^{(t+\Delta)} = \vec{x} + \vec{v}_j \mid \vec{X}^{(t)} = \vec{x}\right) = c_j \cdot \#\text{reactant combinations} \cdot \Delta.$$

Thus, for a fixed combination of reactants, the infinitesimal transition probability is $c_j \cdot \Delta$. Note that we assume here that all change vectors are different, but the above equation is easily extended to the case where change vectors can be identical.

Stochastic Reaction Rate Constant. The above constant c_j is called *stochastic reaction rate constant* and its existence is guaranteed by physical theory [Gil92]. Consider a bimolecular reaction. The value $c_j \cdot \Delta$ is the probability that a fixed pair of reactants collide and that they then undergo a reaction (because a sufficient amount of activation energy is available). For a unimolecular reaction (only one reactant), it is the probability that a fixed reactant molecule undergoes the reaction. The constant c_j depends on the microphysical properties of the reactant species, but also on the temperature and the volume, which may change in time. In the sequel, we assume for simplicity that the temperature and the volume of the system are fixed⁵. Thus, the infinitesimal transition probability

$$P\left(\vec{X}^{(t+\Delta)} = \vec{x} + \vec{v}_j \mid \vec{X}^{(t)} = \vec{x}\right) = c_j \cdot \#\text{reactant combinations} \cdot \Delta.$$

depends on

- the reaction rate constant c_j of reaction R_j ,
- the current state \vec{x} , which gives the number of reactant combinations,
- the time step Δ (but not on the absolute time t).

Let α_j be the function that calculates the product $c_j \cdot \#\text{reactant combinations}$, that is, for a state $\vec{x} = (x_1, \dots, x_n)$ and a reaction R_j of the form

- $S_i + S_k \rightarrow \text{products}$, $i \neq k$, we have $\alpha_j(\vec{x}) = c_j \cdot x_i \cdot x_k$,
- $S_i \rightarrow \text{products}$, we have $\alpha_j(\vec{x}) = c_j \cdot x_i$,

⁵Most concepts that we develop in the sequel will be easily extendable to time-dependent stochastic reaction rate constants.

- $2S_i \rightarrow \text{products}$, we have $\alpha_j(\vec{x}) = c_j \cdot x_i \cdot (x_i - 1) \cdot 0.5$,
- $\emptyset \rightarrow \text{products}$, we have $\alpha_j(\vec{x}) = c_j$.

Example 25: GENE EXPRESSION

The propensity functions $\alpha_1, \dots, \alpha_4$ of the four reactions of the gene expression example are such that for a state $\vec{x} = (x_1, x_2)$ (where x_1 refers to the number of mRNA molecules and x_2 to the number of proteins),

$$\begin{array}{llll}
 R_1 : & \emptyset & \rightarrow & \text{mRNA} & \alpha_1(x_1, x_2) = c_1, \\
 R_2 : & \text{mRNA} & \rightarrow & \text{mRNA} + \text{protein} & \alpha_2(x_1, x_2) = c_2 \cdot x_1, \\
 R_3 : & \text{mRNA} & \rightarrow & \emptyset & \alpha_3(x_1, x_2) = c_3 \cdot x_1, \\
 R_4 : & \text{protein} & \rightarrow & \emptyset & \alpha_4(x_1, x_2) = c_4 \cdot x_2.
 \end{array}$$

The stochastic process $(\vec{X}^{(t)})_{t \geq 0}$ is called a *Markov chain*, because of the Markov property (“Markov”) and the fact that the state space is discrete (“chain”)⁶. Here, time is continuous and continuous-time Markov chains should not be confused with discrete-time Markov chains, where $t = 0, 1, \dots$. Moreover, X is time-homogeneous, which means that the infinitesimal transition probabilities are the same for all time instances t , i.e. they only depend on the length Δ of the time interval but not on the current time instant t .

1.5.2 Chemical Master Equation

Our aim is now to derive an expression for the probabilities

$$P(\vec{X}^{(t)} = \vec{x}) = \sum_{\vec{y} \in \mathbb{Z}_+^N} P(\vec{X}^{(t)} = \vec{x} \mid \vec{X}^{(0)} = \vec{y}) \cdot P(\vec{X}^{(0)} = \vec{y})$$

where the initial values $P(\vec{X}^{(0)} = \vec{y}), y \in \mathbb{Z}_+^N$ are given, as well as the propensity functions $\alpha_1, \dots, \alpha_M$. For an infinitesimal time step of length Δ , we have

$$\begin{aligned}
 P(\vec{X}^{(t+\Delta)} = \vec{x}) &= \overbrace{P(\vec{X}^{(t+\Delta)} = \vec{x} \mid \vec{X}^{(t)} = \vec{x})}^{\text{stayed in } \vec{x} \text{ (no reaction)}} \cdot P(\vec{X}^{(t)} = \vec{x}) \\
 &+ \sum_{\substack{j=1, \\ \vec{x} - \vec{v}_j \geq 0}}^M \underbrace{P(\vec{X}^{(t+\Delta)} = \vec{x} \mid \vec{X}^{(t)} = \vec{x} - \vec{v}_j)}_{\text{in } \vec{x} - \vec{v}_j \text{ at time } t \text{ and reaction } R_j \text{ occurred}} \cdot P(\vec{X}^{(t)} = \vec{x} - \vec{v}_j),
 \end{aligned}$$

⁶Sometimes the term “Markov process” is used.

which is equivalent to

$$\begin{aligned}
 P(\vec{X}^{(t+\Delta)} = \vec{x}) &= \overbrace{\left(1 - \sum_{j=1}^M \alpha_j(\vec{x}) \cdot \Delta\right)}^{\text{stayed in } \vec{x} \text{ (no reaction)}} \cdot P(\vec{X}^{(t)} = \vec{x}) \\
 &+ \sum_{\substack{j=1, \\ \vec{x}-\vec{v}_j \geq 0}}^M \underbrace{\alpha_j(\vec{x} - \vec{v}_j) \cdot \Delta \cdot P(\vec{X}^{(t)} = \vec{x} - \vec{v}_j)}_{\text{in } \vec{x} - \vec{v}_j \text{ at time } t \text{ and reaction } R_j \text{ occurred}}.
 \end{aligned}$$

Thus, we obtain

$$\begin{aligned}
 \frac{d}{dt}P(\vec{X}^{(t)} = \vec{x}) &= \lim_{\Delta \rightarrow 0} \frac{P(\vec{X}^{(t+\Delta)} = \vec{x}) - P(\vec{X}^{(t)} = \vec{x})}{\Delta} \\
 &= - \sum_{j=1}^M \alpha_j(\vec{x}) \cdot P(\vec{X}^{(t)} = \vec{x}) \\
 &+ \sum_{\substack{j=1, \\ \vec{x}-\vec{v}_j \geq 0}}^M \alpha_j(\vec{x} - \vec{v}_j) \cdot P(\vec{X}^{(t)} = \vec{x} - \vec{v}_j).
 \end{aligned} \tag{1.6}$$

The differential equation above is called the *Chemical Master Equation (CME)*. Note that we get an equation for every state $\vec{x} \in \mathbb{Z}_+^N$ and these differential equations are coupled since the derivative on the right side requires the probability $P(\vec{X}^{(t)} = \vec{x} - \vec{v}_j)$ of state $\vec{y} = \vec{x} - \vec{v}_j$.

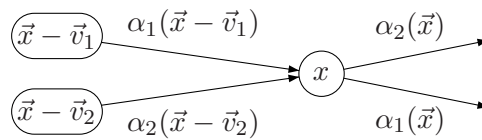


Figure 1.16: Inflow and outflow of probability.

If we combine, for a particular reaction system, the chemical master equation for every state then we have a system of coupled *linear* ODEs where the variables are the probabilities to be in each state of the system, $P(\vec{X}^{(t)} = \vec{x})$. Given an initial probability distribution, the solution of the CME are the probabilities $P(\vec{X}^{(t)} = \vec{x})$ for all states x . In general, the CME may not have a unique solution. The reason is that the differences of the jump times of the process become smaller and smaller and the system “explodes”⁷. Since such pathological cases are only of theoretical interest, we assume that all systems that we consider in the sequel have a unique solution. Moreover, as long as the propensities are of the form described above, the CME will always have a unique solution.

The intuitive meaning of the CME is that the derivative of the probability of state x is the difference between the “inflow of probability” and the “outflow of probability” per time unit. The states are seen as nodes in a flow network and their probability is the amount of fluid, which moves through the network according to the propensities (see Fig. 1.16).

⁷This may happen, if, for instance, the propensities grow exponentially in the state variables.

1.5.3 Expected Populations

Consider the expectation $E(\vec{X}^{(t)})$ of the random vector $\vec{X}^{(t)}$. The i -th entry denotes the expected number of molecules of type S_i at time t . Recall that

$$E(\vec{X}^{(t)}) = \sum_{\vec{x} \in \mathbb{Z}_+^N} \vec{x} \cdot P(\vec{X}^{(t)} = \vec{x})$$

Fig. 1.17 illustrates the solution of the CME for the enzyme reaction (see Ex. 1) with initial condition $P(\vec{X}^{(0)} = (10, 10, 0, 0))$.

In the sequel, we will exploit the relationship

$$\begin{aligned} \frac{d}{dt} E(f(\vec{X}^{(t)})) &= \sum_{\vec{x} \in \mathbb{Z}_+^N} f(\vec{x}) \cdot \frac{d}{dt} P(\vec{X}^{(t)} = \vec{x}) \\ &= \sum_{j=1}^M E(\alpha_j(\vec{X}^{(t)}) \cdot (f(\vec{X}^{(t)} + \vec{v}_j) - f(\vec{X}^{(t)}))), \end{aligned} \quad (1.7)$$

where $f : \mathbb{Z}_+^N \rightarrow \mathbb{R}^N$ is a function that is independent of t . As a special case, we obtain

$$\frac{d}{dt} E(\vec{X}^{(t)}) = \sum_{j=1}^M \vec{v}_j \cdot E(\alpha_j(\vec{X}^{(t)})), \quad (1.8)$$

which means that each reaction contributes to the change of the expectation with the product of the expected propensity and the change vector. The expected propensity of reaction R_j is given by

$$E(\alpha_j(\vec{X}^{(t)})) = \begin{cases} c_j & \text{if } \alpha_j(\vec{x}) = c_j, \\ c_j \cdot E(X_i^{(t)}) & \text{if } \alpha_j(\vec{x}) = c_j \cdot x_i, \\ c_j \cdot E(X_i^{(t)} \cdot X_k^{(t)}) & \text{if } \alpha_j(\vec{x}) = c_j \cdot x_i \cdot x_k, \\ \frac{1}{2} c_j \cdot (E((X_i^{(t)})^2) - E(X_i^{(t)})) & \text{if } \alpha_j(\vec{x}) = \frac{1}{2} c_j \cdot (x_i^2 - x_i). \end{cases}$$

Since we get a new equation for all expectations involving the product of two random variables (third and fourth case, i.e., whenever α_j is not linear in the elements of \vec{x}), which again involve expectations for which we get new equations, this leads to an infinite series of differential equations. However, if all propensity functions α_j are linear (first and second case, i.e., at most one reactant molecule), the equation can be simplified to

$$\frac{d}{dt} E(\vec{X}^{(t)}) = \sum_{j=1}^M \vec{v}_j \cdot \alpha_j(E(\vec{X}^{(t)})),$$

where we used the second property of the expectation (see page 19). This gives a system of differential equations, which is identical to the reaction rate equations except that we consider populations instead of concentrations. In general, however, the solution of the reaction rate equations is different from the average populations of the Markov chain.

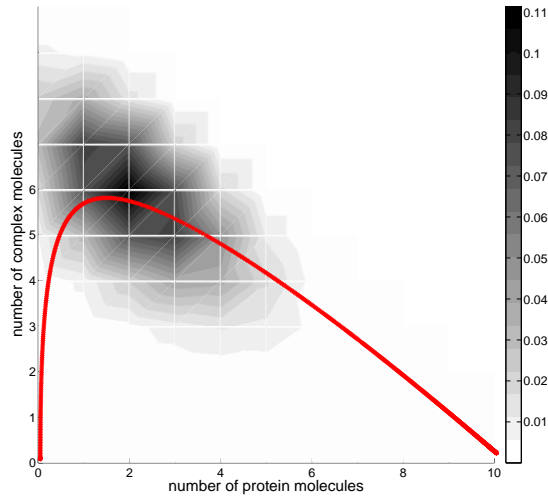


Figure 1.17: Solution of the CME for the enzyme reaction at time $t = 0.44$. Expected populations of P and C molecules are depicted in red.

Example 26: EXPECTATION OF THE GENE EXPRESSION EXAMPLE

Consider the gene expression example (see Ex. 23). Assume that $E(\vec{X}^{(0)}) = (0, 0)$. We compute $E(\vec{X}^{(t)})$ for $t > 0$ by solving

$$\begin{aligned}
 \frac{d}{dt}E(\vec{X}^{(t)}) &= \sum_{j=1}^M \vec{v}_j \cdot E(\alpha_j(\vec{X}^{(t)})) \\
 &= \sum_{j=1}^4 \vec{v}_j \cdot \alpha_j(E(\vec{X}^{(t)})) \\
 &= (1, 0) \cdot c_1 + (0, 1) \cdot c_2 \cdot E(X_1^{(t)}) \\
 &\quad + (-1, 0) \cdot c_3 \cdot E(X_1^{(t)}) + (0, -1) \cdot c_4 \cdot E(X_2^{(t)}).
 \end{aligned}$$

The ODE can be simplified to

$$\begin{aligned}
 \frac{d}{dt}E(X_1^{(t)}) &= c_1 - c_3 \cdot E(X_1^{(t)}) \\
 \frac{d}{dt}E(X_2^{(t)}) &= c_2 \cdot E(X_1^{(t)}) - c_4 \cdot E(X_2^{(t)}).
 \end{aligned}$$

Let us now convert Eq. 1.8 from populations to concentrations. Assume that $V \in \mathbb{R}_{>0}$ is the volume (in liters). Choosing $f(\vec{x}) = \vec{x} \cdot V^{-1}$ in Eq. 1.7

yields

$$\begin{aligned} \frac{d}{dt}E(f(\vec{X}(t))) &= \sum_{j=1}^M E(\alpha_j(\vec{X}(t)) \cdot (f(\vec{X}(t) + \vec{v}_j) - f(\vec{X}(t)))) \\ &= \sum_{j=1}^M \vec{v}_j \cdot E(V^{-1} \cdot \alpha_j(\vec{X}(t))). \end{aligned}$$

If $\vec{Y}(t) := f(\vec{X}(t)) = \vec{X}(t) \cdot V^{-1}$ represents the concentrations of all species at time t ,

$$\frac{d}{dt}E(\vec{Y}(t)) = \sum_{j=1}^M \vec{v}_j \cdot E(\underbrace{V^{-1} \cdot \alpha_j(V \cdot \vec{Y}(t))}_{=: \beta_j(\vec{Y}(t))}). \quad (1.9)$$

For $j \in \{1, 2, \dots, M\}$, $\vec{y} = (y_1, \dots, y_N) \in \mathbb{R}_+^N$, let $\beta_j : \mathbb{R}_+^N \rightarrow \mathbb{R}_+$ be the function with $V \cdot \beta_j(\vec{y}) = \alpha_j(V \cdot \vec{y})$, that is⁸,

$$\beta_j(\vec{y}) = V^{-1} \cdot \alpha_j(V \cdot \vec{y}) = \begin{cases} V^{-1} \cdot c_j & \text{if } \alpha_j(\vec{x}) = c_j, \\ c_j \cdot y_i & \text{if } \alpha_j(\vec{x}) = c_j \cdot x_i, \\ V \cdot c_j \cdot y_i \cdot y_k & \text{if } \alpha_j(\vec{x}) = c_j \cdot x_i \cdot x_k, \\ \frac{1}{2}c_j \cdot y_i \cdot (V \cdot y_i - 1) & \text{if } \alpha_j(\vec{x}) = \frac{1}{2}c_j \cdot x_i \cdot (x_i - 1). \end{cases}$$

As before, we can simplify Eq. 1.9 to

$$\frac{d}{dt}E(\vec{Y}(t)) = \sum_{j=1}^M \vec{v}_j \cdot \beta_j(E(\vec{Y}(t))), \quad (1.10)$$

which equals the reaction rate equations with reaction rate constants

$$k_j = \begin{cases} V^{-1} \cdot c_j & \text{if } \alpha_j(\vec{x}) = c_j, \\ c_j & \text{if } \alpha_j(\vec{x}) = c_j \cdot x_i. \end{cases}$$

In the general case, however, (not all functions β_j are necessarily linear) it can be shown that the ODE

$$\frac{d}{dt}\vec{z}(t) = \sum_{j=1}^M \vec{v}_j \cdot \beta_j(\vec{z}(t))$$

approximates the values $E(\vec{Y}(t))$ (with initial condition $\vec{z}(0) = E(\vec{Y}(0))$). As molecule numbers and volume approach infinity this approximation is shown to be exact [Kur72]. Note that for reactions of the form $2S_i \rightarrow \dots$, we approximate $\beta_j(\vec{y}) = \frac{1}{2}c_j \cdot y_i \cdot (V \cdot y_i - 1) \approx \frac{1}{2}c_j \cdot V \cdot y_i^2$ since $V \cdot y_i - 1 \approx V \cdot y_i$ if the population of S_i is large.

⁸Note that we can safely extend the domain of α_j from \mathbb{Z}_+^N to \mathbb{R}_+^N .

1.6 Summary

We have discussed two approaches for the analysis of biochemical reaction networks. For the approach based on reaction rate equations we assume that the concentrations of the chemical species change continuously and deterministically in time. This gives us a system of (non-linear) differential equations, whose solution are the concentrations of the species as a function in time. The approach using the chemical master equation is based on a stochastic description that considers populations of chemical species that change at discrete points in time. Our motivation for the latter approach is that even in genetically identical cell populations with equal amounts of cellular components, the outcomes of certain experiments can be different due to intrinsic noise.

The stochastic model assumes random collisions between molecules that may change the populations of the species. The populations are represented by the random vector $\vec{X}^{(t)}$ whose i -th entry represents the number of molecules of type S_i at time t . Given an initial population \vec{x}_0 and a propensity function α_j for each reaction R_j , $j \in \{1, \dots, M\}$, the probabilities $P(\vec{X}^{(t)} = \vec{x})$ are given by the solution of the chemical master equation (see Eq. 1.6). The state space of the stochastic model can be extremely large, e.g., if we start in the case of the enzyme example (Ex. 1) with 500 enzymes and 500 substrates, the number of reachable states is 125751. Thus, the CME would have 125751 equations in this case. For other examples, such as the gene expression example (see Ex. 23), the number of reachable states is even infinite. If the populations of the chemical species are large, we can approximate the expected concentrations with the solution of the reaction rate equations. For systems, where certain species occur with small populations, we can resort to *stochastic simulation*, which is the topic of the next chapter.

Chapter 2

Stochastic Simulation

2.1 Introduction

Since a numerical solution of the chemical master equation is in most cases computationally very expensive or even infeasible, in this chapter we present an alternative way to analyse Markov chains that describe networks of biochemical reactions. The idea is to generate trajectories of the Markov chain by using pseudo random numbers. Once enough trajectories are generated, we use arguments from statistics to estimate probabilities of certain events. In Markov chain theory, the generation of trajectories is called *Monte-Carlo simulation*. For coupled chemical reactions, however, this approach is usually referred to as *stochastic simulation* or *Gillespie simulation*.

2.2 Trajectory Generation

Consider a Markov chain with state space \mathbb{Z}_+^N that describes the dynamics of a network of M biochemical reactions with N molecular species and assume that x_0 is the initial state of the system¹. In order to generate a prefix of a trajectory $t \mapsto \vec{X}^{(t)}(\omega)$, that is, values $X^{(t)}(\omega)$ during a certain interval $[0, h]$, we need to know

- a) the distribution of the time that the process remains in a state once that state is entered,
- b) the probability that, for $j \in \{1, 2, \dots, M\}$, if the current state is x , the next state is $x + v_j$, which means that reaction R_j occurred.

The random variables that correspond to the distributions mentioned in case a) are called *residence times*. We refer as *jump probabilities* to the probabilities mentioned in case b).

¹We omit the vector notation in the sequel and simply write x instead of \vec{x} and X instead of \vec{X} , etc.

2.2.1 Residence Times

For the continuous-time Markov chain $(X^{(t)})_{t \geq 0}$ on a sample space Ω we define *jump times* $T_0(\omega) = 0$ and, for $n = 0, 1, \dots$,

$$T_{n+1}(\omega) = \inf\{t \geq T_n(\omega) \mid X^{(t)}(\omega) \neq X^{(T_n)}(\omega)\},$$

where $\omega \in \Omega$ and $T_{n+1}(\omega) = \infty$ if $X^{(t)}(\omega) = X^{(T_n)}(\omega)$ for all $t > T_n(\omega)$. Note that T_n is a random variable with $T_n : \Omega \rightarrow \mathbb{R}_{\geq 0} \cup \{\infty\}$.

Next we define the sequence of residence times of X as $D_0(\omega) = 0$ for all $\omega \in \Omega$ and, for $n = 0, 1, \dots$,

$$D_{n+1} = T_{n+1} - T_n,$$

where we assume that $\infty - x = \infty$ if $x \in \mathbb{R}$.

After n jumps, process X is in state $X^{(T_n)}$ and remains there for D_{n+1} time units.

Let us now consider the probability to remain longer than s time units in state x . It holds that

$$P(D_{n+1} > s \mid X^{(T_n)} = x) = P(D_1 > s \mid X^{(0)} = x),$$

since the transition probabilities of X are independent of t . For an infinitesimal time interval $[s, s + \Delta)$,

$$\begin{aligned} & P(D_1 > s + \Delta \mid X^{(0)} = x) \\ &= P(D_1 > s + \Delta, D_1 > s \mid X^{(0)} = x) \\ &= P(D_1 > s + \Delta \mid D_1 > s, X^{(0)} = x) \cdot P(D_1 > s \mid X^{(0)} = x) \\ &= \underbrace{\left(1 - \sum_{j=1}^M \alpha_j(x) \cdot \Delta\right)}_{\text{no reaction within } [0, \Delta]} \cdot P(D_1 > s \mid X^{(0)} = x). \end{aligned}$$

Note that we used the Markov property in the last step of the above equation. The probability of no reaction within $[0, \Delta]$ is independent of the fact that the process already remained s time units in state x . For the derivation we

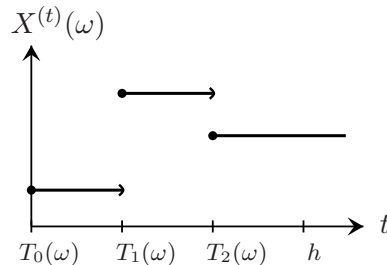


Figure 2.1: Jump times T_0, T_1, \dots

get

$$\begin{aligned}
 & \frac{d}{ds} P(D_1 > s \mid X^{(0)} = x) \\
 &= \lim_{\Delta \rightarrow 0} \frac{P(D_1 > s + \Delta \mid X^{(0)} = x) - P(D_1 > s \mid X^{(0)} = x)}{\Delta} \\
 &= - \underbrace{\sum_{j=1}^M \alpha_j(x)}_{=: \alpha_0(x)} \cdot P(D_1 > s \mid X^{(0)} = x).
 \end{aligned}$$

Since $P(D_1 > 0 \mid X^{(0)} = x) = 1$, the differential equation has the solution

$$P(D_1 > s \mid X^{(0)} = x) = e^{-\alpha_0(x) \cdot s},$$

which means that the distribution of the residence time in state x is negative exponential with parameter

$$\sum_{j=1}^M \alpha_j(x) = \alpha_0(x).$$

Note that the exponential distribution has the memoryless property, that is, for an exponentially distributed random variable D , $t, s \in \mathbb{R}_{>0}$

$$P(D \leq s + t \mid D > t) = P(D \leq s).$$

An alternative way to prove that the residence times of X are exponentially distributed is to derive the memoryless property of the residence times from the Markov property. Then one can exploit the fact that the exponential distribution is the only continuous distribution having the memoryless property.

2.2.2 Number of Jumps

Let us now consider the distribution

$$P(A(t) = l \mid X^{(0)} = x), \quad l = 0, 1, \dots$$

of the random variable $A(t)$, which represents the number of jumps of X within $[0, t]$. For instance, in Fig. 2.1, $A(h, \omega) = 2$. For $l = 0$, we get

$$\begin{aligned}
 P(A(t) = 0 \mid X^{(0)} = x) &= P(D_1 > t \mid X^{(0)} = x) = e^{-\alpha_0(x)t} \\
 &= 1 - \alpha_0(x)t + \frac{(\alpha_0(x)t)^2}{2!} - \frac{(\alpha_0(x)t)^3}{3!} + \dots \\
 &= 1 - \alpha_0(x)t + o(t),
 \end{aligned}$$

where $o(t)$ is a function such that $o(0) = 0$ and $\lim_{t \rightarrow 0} o(t)/t = 0$. For $l \geq 1$, we get

$$\begin{aligned}
 P(A(t) \geq 1 \mid X^{(0)} = x) &= 1 - P(A(t) = 0 \mid X^{(0)} = x) \\
 &= \alpha_0(x)t - o(t).
 \end{aligned}$$

In a similar way, it can be shown that² $P(A(t) \geq 2 \mid X^{(0)} = x) = o(t)$. Thus,

$$\begin{aligned} P(A(t) = 1 \mid X^{(0)} = x) &= 1 - P(A(t) = 0 \mid X^{(0)} = x) \\ &\quad - P(A(t) \geq 2 \mid X^{(0)} = x) \\ &= \alpha_0(x)t + o(t). \end{aligned}$$

2.2.3 Jump Probabilities

With the above preparation, we can now consider the probability that, if the current state is x , the next state is $x + v_j$, that is, reaction R_j occurred. For this we define for the given Markov chain X the embedded process $(Z_n)_{n \in \mathbb{Z}_+}$ by

$$Z_n = X^{(T_n)}.$$

Thus, if on the trajectory $X^{(t)}(\omega)$ the process visits the states $X^{(T_0(\omega))}(\omega) = x_0$, $X^{(T_1(\omega))}(\omega) = x_1, \dots$, then $Z_0(\omega) = x_0, Z_1(\omega) = x_1, \dots$.

The stochastic process Z has the Markov property since for all $n, k \in \mathbb{Z}_+$, $x, y, x_0, x_1, \dots, x_{n-1} \in \mathbb{Z}_+^N$,

$$\begin{aligned} &P(Z_{n+k} = y \mid Z_n = x, Z_{n-1} = x_{n-1}, \dots, Z_0 = x_0) \\ &= P(X^{(T_{n+k})} = y \mid X^{(T_n)} = x, X^{(T_{n-1})} = x_{n-1}, \dots, X^{(T_0)} = x_0) \\ &= P(X^{(T_{n+k})} = y \mid X^{(T_n)} = x) \\ &= P(Z_{n+k} = y \mid Z_n = x), \end{aligned}$$

where we used the Markov property of X . Therefore, $(Z_n)_{n \in \mathbb{Z}_+}$ is called discrete-time Markov chain. Note that time is discrete for Z because $n \in \mathbb{Z}_+$. Next we are interested in the transition probabilities $P(Z_{n+1} = y \mid Z_n = x)$. Obviously, they do not depend on n since the infinitesimal transition probabilities of X do not depend on t . More precisely,

$$\begin{aligned} &P(Z_{n+1} = y \mid Z_n = x) = P(X^{(T_{n+1})} = y \mid X^{(T_n)} = x) \\ &= P(X^{(T_1)} = y \mid X^{(T_0)} = x) = P(Z_1 = y \mid Z_0 = x). \end{aligned}$$

Now consider, for a state x with $\alpha_0(x) > 0$,

$$\begin{aligned} &P(X^{(t+\Delta)} = x + v_j \mid X^{(t)} = x) = P(X^{(\Delta)} = x + v_j \mid X^{(0)} = x) \\ &= P(Z_1 = x + v_j, A(\Delta) = 1 \mid Z_0 = x) \\ &= P(Z_1 = x + v_j \mid A(\Delta) = 1, Z_0 = x) \cdot P(A(\Delta) = 1 \mid X^{(0)} = x) \\ &= P(Z_1 = x + v_j \mid Z_0 = x) \cdot P(A(\Delta) = 1 \mid X^{(0)} = x). \end{aligned}$$

²We write $o(t)$ for any function with $o(0) = 0$ and $\lim_{t \rightarrow 0} o(t)/t = 0$.

We divide by Δ and get, for $\Delta \rightarrow 0$,

$$\begin{aligned} & \lim_{\Delta \rightarrow 0} (1/\Delta \cdot P(X^{t+\Delta} = x + v_j \mid X^t = x)) \\ &= P(Z_1 = x + v_j \mid Z_0 = x) \cdot \lim_{\Delta \rightarrow 0} (1/\Delta \cdot P(A(\Delta) = 1 \mid X^{(0)} = x)). \end{aligned}$$

From Section 1.5.1 we know that

$$P(X^{t+\Delta} = x + v_j \mid X^t = x) = \alpha_j(x) \cdot \Delta + o(\Delta),$$

where $o(\Delta)$ accounts for the probability of more than one reaction during $[0, \Delta)$. For the number of jumps we found that

$$P(A(\Delta) = 1 \mid X^{(0)} = x) = \alpha_0(x) \cdot \Delta + o(\Delta).$$

Thus, the combination yields

$$\lim_{\Delta \rightarrow 0} \frac{\alpha_j(x) \cdot \Delta + o(\Delta)}{\Delta} = P(Z_1 = x + v_j \mid Z_0 = x) \cdot \lim_{\Delta \rightarrow 0} \frac{\alpha_0(x) \cdot \Delta + o(\Delta)}{\Delta}.$$

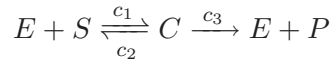
Since $\lim_{\Delta \rightarrow 0} \frac{o(\Delta)}{\Delta} = 0$ we get

$$\begin{aligned} \alpha_j(x) &= P(Z_1 = x + v_j \mid Z_0 = x) \cdot \alpha_0(x) \\ \iff P(Z_1 = x + v_j \mid Z_0 = x) &= \frac{\alpha_j(x)}{\alpha_0(x)}. \end{aligned}$$

This means that, in state x , reaction R_j is selected with a probability equal to its fraction $\alpha_j(x)$ of the total propensity $\alpha_0(x)$. For the case $\alpha_0(x) = 0$ there are no reactions possible in state x . So if $\alpha_0(x) = 0$ then $P(Z_1 = x \mid Z_0 = x) = 1$ and $P(Z_1 = y \mid Z_0 = x) = 0$ for states $y \neq x$.

Example 27: ENZYME REACTION

Consider the enzyme reaction



and assume $c_1 = c_2 = c_3 = 1$. If $x = (3, 2, 1, 0)$, that is, we have three enzyme molecules, two substrates, one complex molecule, and zero proteins, then reaction $E + S \rightarrow C$ occurs with probability

$$P(Z_1 = (2, 1, 2, 0) \mid Z_0 = x) = \frac{\alpha_1(x)}{\alpha_0(x)} = \frac{3 \cdot 2}{8},$$

reaction $C \rightarrow E + S$ occurs with probability

$$P(Z_1 = (4, 3, 0, 0) \mid Z_0 = x) = \frac{\alpha_2(x)}{\alpha_0(x)} = \frac{1}{8},$$

and reaction $C \rightarrow E + P$ occurs with probability

$$P(Z_1 = (4, 2, 0, 1) \mid Z_0 = x) = \frac{\alpha_3(x)}{\alpha_0(x)} = \frac{1}{8}.$$

2.2.4 Algorithm

The algorithm that is used to generate a prefix of a trajectory $t \mapsto \vec{X}^{(t)}(\omega)$ takes as input

- the time horizon h of interest,
- propensity functions $\alpha_1, \dots, \alpha_M$,
- change vectors v_1, \dots, v_M ,
- an initial state $x_0 \in \mathbb{Z}_+^N$.

It performs the following steps:

1. Initialize time $t := 0$ and state $x := x_0$.
2. Store x and t .
3. Draw an exponentially distributed pseudo random number τ for parameter $-\alpha_0(x) = -\sum_{j=1}^M \alpha_j(x)$ and update $t = t + \tau$.
4. Draw a pseudo random number j according to the discrete distribution

$$\frac{\alpha_1(x)}{\alpha_0(x)}, \dots, \frac{\alpha_M(x)}{\alpha_0(x)}$$

and update $x = x + v_j$.

5. If $t < h$ goto step 2 else stop.

It is important to note that a single trajectory provides only very little information since we do not know whether this is a trajectory that is very likely or not. If, however, a large number of trajectories is produced, we can use arguments from statistics to estimate probabilities, expectations and variances.

Inverse Transform Method. Steps 3 and 4 of the algorithm require the generation of pseudo random numbers that follow a certain distribution. In most software packages, however, one can only produce numbers that are uniformly distributed on the interval $(0, 1)$. The *inverse transform method* allows us to transform a uniformly distributed pseudo random number $u = U(\omega)$ into a pseudo random number $\tau = X(\omega)$ that has cumulative distribution function \tilde{F} .

Assume that U is a random variable that is uniformly distributed on $(0, 1)$. We define

$$X := \inf_{z \in \mathbb{R}} \{\tilde{F}(z) \geq U\}.$$

Recall Def. 5 on page 16 and Def. 12 on page 24. Note that if the inverse function \tilde{F}^{-1} of \tilde{F} exists then $\tilde{F}^{-1}(y) = \inf_{z \in \mathbb{R}} \{\tilde{F}(z) \geq y\}$. But even if the inverse function does not exist, we use the notation $\tilde{F}^{-1}(y)$ for $\inf_{z \in \mathbb{R}} \{\tilde{F}(z) \geq y\}$, since this function has similar properties, e.g. $\tilde{F}(\tilde{F}^{-1}(y)) = y$. An example for the case that no inverse function of \tilde{F} exists is the cumulative distribution function of a random variable that represents the pips of a die. For the cumulative distribution function of X , we compute

$$\begin{aligned}
 P(X \leq z) &= P(\tilde{F}^{-1}(U) \leq z) \\
 &= P(\{\omega \mid \tilde{F}^{-1}(U(\omega)) \leq z\}) \\
 &= P(\{\omega \mid \tilde{F}(\tilde{F}^{-1}(U(\omega))) \leq \tilde{F}(z)\}) \\
 &= P(\{\omega \mid \tilde{U}(\omega) \leq \tilde{F}(z)\}) \\
 &= P(\tilde{U} \leq \tilde{F}(z)) \\
 &= \tilde{F}(z)
 \end{aligned}$$

where we used that \tilde{F} is monotone (third line) and that the cumulative distribution function of U is such that $P(U \leq y) = y$ (last line). For the exponential distribution (with parameter $-\lambda$), this yields

$$\begin{aligned}
 y := \tilde{F}(z) &= 1 - e^{-\lambda z} \\
 \iff 1 - y &= e^{-\lambda z} \\
 \iff \ln(1 - y) &= -\lambda z \\
 \iff z &= -\frac{1}{\lambda} \ln(1 - y) = \tilde{F}^{-1}(y).
 \end{aligned}$$

Thus, if U is uniformly distributed on $(0, 1)$, the random variable $X = \tilde{F}^{-1}(U) = -\frac{1}{\lambda} \ln(1 - U)$ is exponentially distributed with parameter $-\lambda$. Moreover, since $1 - U$ has the same distribution as U , we can set $X = -\frac{1}{\lambda} \ln(U)$. This means step 3 of the algorithm for the trajectory generation consists of the following two tasks:

- a) Draw a pseudo random number u with uniform distribution on $(0, 1)$.
- b) Set $\tau := -\frac{\ln(u)}{\alpha_0(x)}$.

Let \tilde{F} be the cumulative distribution function of a discrete random variable X with $P(X = 1) = p_1, P(X = 2) = p_2, \dots, P(X = M) = p_M$ (see Fig. 2.2 for an illustration). Then

$$\tilde{F}(x) = \begin{cases} \sum_{k=1}^{\min(\lfloor x \rfloor, M)} p_k & \text{if } x \geq 1, \\ 0 & \text{if } x < 1. \end{cases}$$

2.3. OUTPUT ANALYSIS

For $u \in (0, 1)$, the inverse is given by

$$\tilde{F}^{-1}(u) = \min\{i \in \{1, \dots, M\} \mid \sum_{k=1}^i p_k \geq u\}.$$

This means step 4 of the algorithm for the trajectory generation consists of the following two tasks:

- a) Draw a pseudo random number u with uniform distribution on $(0, 1)$.
- b) Set $j := \min\{i \in \{1, \dots, M\} \mid \alpha_1(x) + \dots + \alpha_i(x) \geq u \cdot \alpha_0(x)\}$.

The above considerations can be used to write a simple computer program to generate trajectories of the Markov chain X . The only requirement is the availability of a method that generates pseudo random variables that are uniformly distributed on the interval $(0, 1)$. In the sequel, we refer to a generated trajectory as (*simulation*) *run* and to the generation of a number of runs as Gillespie simulation.

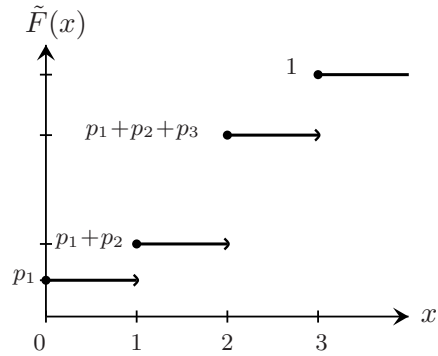


Figure 2.2: Function \tilde{F} for discrete RV with $M = 4$, $p_4 = 1 - (p_1 + p_2 + p_3)$.

2.3 Output Analysis

Assume that we generate 100 runs of a Markov chain using the algorithm presented in the previous section. For each run we can check whether a certain event occurred or not. E.g. assume that in 67 of the 100 runs we have that $X_i(0.4) = 10$. Then the relative frequency of the event “ten molecules of type i at time 0.4” is $\frac{67}{100}$. Intuitively, we can deduce that $P(X_i(0.4) = 10) \approx \frac{67}{100}$. However, we do not know how good this estimation is since the runs are based on pseudo random numbers. If we do another 100 runs, we could get a very different frequency for the same event. In this section, we discuss how we can estimate probabilities of events and expected values of random variables using Gillespie simulation. We will see how we can determine the quality of an estimated value. For the latter we need arguments from statistics.

2.3.1 Fundamental Results from Statistics

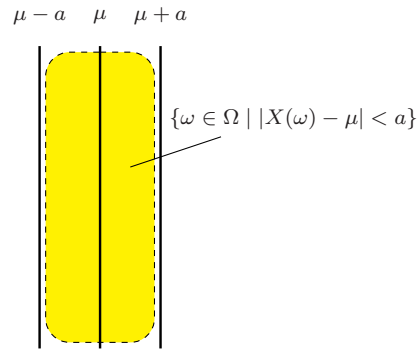
In the sequel we recall some important properties about sums of random variables.

Laws of Large Numbers. Consider n realizations $x_1 = X(\omega_1)$, $x_2 = X(\omega_2)$, \dots , $x_n = X(\omega_n)$ of a random variable X with finite expectation μ . They could, for instance, be generated by repeating an experiment n times under equal conditions. Intuitively, for large n the sample mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \approx \mu.$$

In order to determine the quality of the above approximation, we recall *Chebyshev's Inequality*.

Let Y be a random variable with finite expectation $E(Y)$ and finite variance $VAR(Y)$. Assume that besides $E(Y)$ and $VAR(Y)$, nothing is known about Y (e.g. the cumulative probability distribution). Our aim is to reason about the deviation of Y from its expectation. According to Chebyshev's Inequality (see Fig. 2.3 for an illustration), for any $a > 0$



$$P(|Y - E(Y)| \geq a) \leq \frac{VAR(Y)}{a^2}. \quad (2.1)$$

Figure 2.3: Chebyshev's Inequality.

In order to apply the inequality above, we assume that x_1, x_2, \dots, x_n are realizations of independent random variables X_1, X_2, \dots, X_n , all having the same distribution as X (see also Def. 8). Define

$$Z_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Then $E(Z_n) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \mu$ and, if X has finite variance $VAR(X) = \sigma^2$,

$$VAR(Z_n) = VAR\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n VAR(X_i) = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

(See also page 21 for the properties of the variance operator.) Thus, Eq. (2.1) gives us, for any $\epsilon > 0$,

$$P(|Z_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n \cdot \epsilon^2}.$$

If ϵ is given, we can make $\frac{\sigma^2}{n \cdot \epsilon^2}$ arbitrarily small by increasing n . Thus, for any $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} P(|Z_n - \mu| \geq \epsilon) = 0 \quad (\text{or, equivalently, } \lim_{n \rightarrow \infty} P(|Z_n - \mu| < \epsilon) = 1),$$

which is known as the *weak law of large numbers*³. Here, the sequence $\{Z_n\}_{n \geq 1}$ of random variables converges “weakly” since only the corresponding probabilities converge.

Example 28: BERNOULLI TRIALS

Assume that we repeat an experiment whose outcome can be either “success” or “failure”. Such an experiment is called a Bernoulli trial. For n Bernoulli trials and $1 \leq j \leq n$, let X_j be the random variable that is 1 if the j -th trial is success and 0 otherwise. Moreover, let $P(X_j = 1) = p$ for all j . Then $E(X_1) = E(X_2) = \dots = E(X_n) = p$. According to the weak law of large numbers, for any $\epsilon > 0$,

$$P\left(\left|\frac{1}{n} \sum_{j=1}^n X_j - p\right| < \epsilon\right) \rightarrow 1$$

as $n \rightarrow \infty$.

An even stronger result than the weak law of large numbers provides the *strong law of large numbers*, which states that

$$P\left(\lim_{n \rightarrow \infty} |Z_n - \mu| = 0\right) = 1.$$

Here, we measure the probability of all outcomes $\omega \in \Omega$ with $\lim_{n \rightarrow \infty} |Z_n(\omega) - \mu| = 0$ and find that this event occurs with probability one. The strong law of large numbers implies the weak law of large numbers.

Standardized Random Variables. In the sequel, we consider standardized random variables, that is, for a random variable X with finite $E(X) = \mu$ and finite $VAR(X) = \sigma^2$, we define

$$X^* = \frac{X - \mu}{\sigma}.$$

Then

$$E(X^*) = E\left(\frac{X - \mu}{\sigma}\right) = \frac{E(X) - \mu}{\sigma} = 0$$

and (see also page 21)

$$VAR(X^*) = VAR\left(\frac{X - \mu}{\sigma}\right) = \frac{VAR(X)}{\sigma^2} = 1.$$

In the special case of Z_n as defined above, we get

$$Z_n^* = \frac{Z_n - E(Z_n)}{\sqrt{VAR(Z_n)}} = \frac{Z_n - \mu}{\sigma/\sqrt{n}}.$$

³Note that, as opposed to the strong law of large numbers, for the weak law of large numbers pairwise independence of X_1, \dots, X_n is already sufficient and (full) independence is not necessary.

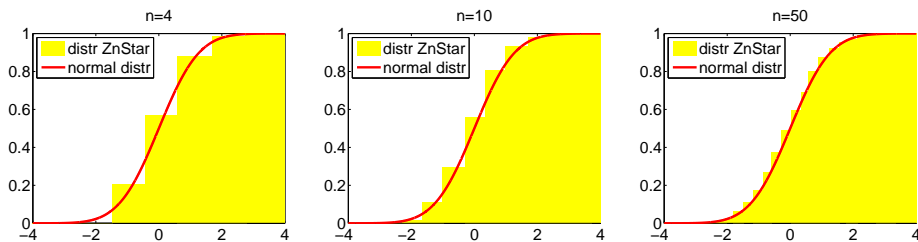


Figure 2.4: The distribution of Z_n^* approaches the standard normal distribution as n increases.

Example 29: BERNOULLI TRIALS

Consider again the Bernoulli experiments as defined in Ex. 28. Fig. 2.4 shows a plot of the cumulative probability distribution of Z_n^* (in yellow) for different values of n and the standard normal distribution $\mathcal{N}(0, 1)$ (in red). It can be seen that the distribution of Z_n^* approaches $\mathcal{N}(0, 1)$ as n increases.

Central Limit Theorem. Recall that $Z_n = \frac{1}{n} \sum_{i=1}^n X_i$, where the X_i are independent and identically distributed random variables with finite expectation μ and finite variance $\sigma^2 > 0$. According to the central limit theorem, the distribution of $Z_n^* = \frac{Z_n - \mu}{\sigma/\sqrt{n}}$ converges to the normal distribution with expectation 0 and variance 1 (standard normal distribution).

The standard normal distribution has density

$$\varphi_{0,1}(t) = \frac{1}{\sqrt{2\pi}} e^{-0.5t^2},$$

where $t \in \mathbb{R}$. Thus, for $x, y \in \mathbb{R}$, $x < y$,

$$\lim_{n \rightarrow \infty} P(x < Z_n^* < y) = \frac{1}{\sqrt{2\pi}} \int_x^y e^{-0.5t^2} dt.$$

Note that if a random variable X is normally distributed with expectation $\hat{\mu}$ and variance $\hat{\sigma}^2$, written $X \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$, then, for $a, b \in \mathbb{R}$, $aX + b$ has distribution $\mathcal{N}(a\hat{\mu} + b, (a\hat{\sigma})^2)$. Therefore, Z_n is approximately normally distributed with expectation μ and variance σ^2/n , since $Z_n = \sigma/\sqrt{n} \cdot Z_n^* + \mu$ and $Z_n^* \sim \mathcal{N}(0, 1)$.

In the sequel we will exploit the central limit theorem to determine how close the sample mean Z_n is to μ .

2.3.2 Interval Estimation

Let Y be a Markov chain that represents a chemical reaction network. Assume that we are interested in the probability of a certain event A , such as

$Y_i(t) = x$ (“ x molecules of type i at time t ”). For $\omega \in \Omega$, define

$$\chi_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A, \end{cases}$$

Then $E(\chi_A) = 1 \cdot P(\chi_A = 1) + 0 \cdot P(\chi_A = 0) = P(\chi_A = 1) = P(A)$. Thus, an estimate of $E(\chi_A)$ is also an estimate of $P(A)$. In the sequel, we concentrate on estimates of the expectation of some random variable. Of course, the procedure described below can also be used for expectations of the form $E(Y_i(t))$, that is, the expected number of molecules of type i at time t .

Let X be a random variable and let X_1, X_2, \dots, X_n be independent random variables, all having the same distribution as X . For instance, if we generate n trajectories of Y using Gillespie simulation, the random variable X_j represents the value of χ_A in the j -th run (i.e., it is either 1 or 0) or the number of molecules at time t in the j -th run.

Let $\mu = E(X) = E(X_1) = \dots = E(X_n)$. We discuss the construction of interval bounds I_l, I_r such that “with high probability” $\mu \in [I_l, I_r]$. Since these bounds depend on X_1, X_2, \dots, X_n , I_l and I_r are random variables.

Estimators for Expectation and Variance. The sample mean

$$Z_n = \frac{1}{n} \sum_{i=1}^n X_i$$

is called an *estimator* for μ . Above, we calculated that $E(Z_n) = \mu$, which means that Z_n is an unbiased estimator. Let $\text{VAR}(X) = \text{VAR}(X_1) = \dots = \text{VAR}(X_n) = \sigma^2$ and recall that $\text{VAR}(Z_n) = \frac{\sigma^2}{n}$. According to the central limit theorem,

$$Z_n^* = \frac{Z_n - \mu}{\sigma/\sqrt{n}}$$

approximately follows a standard normal distribution. We are interested in the deviation $|Z_n - \mu|$ of the estimator Z_n and μ , that is, in

$$P(|Z_n^*| \leq z) = P\left(\frac{|Z_n - \mu|}{\sigma/\sqrt{n}} \leq z\right) = P(|Z_n - \mu| \leq z \cdot \sigma/\sqrt{n}). \quad (2.2)$$

Since we do not know σ (or σ^2) we estimate σ^2 with the sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - Z_n)^2.$$

Note that S^2 is an unbiased estimator for σ^2 and that

$$S^2 = \frac{\sum_{i=1}^n X_i^2}{n-1} - \frac{(\sum_{i=1}^n X_i)^2}{(n-1)n}. \quad (2.3)$$

Confidence Level. In order to control the estimation error, we want the constraint in Eq. 2.2 to hold with a high probability. We choose a *confidence level* β and define z such that $\beta = P(|Z_n^*| \leq z)$. Usually, $\beta \in \{0.95, 0.99\}$ and an approximation for z can be found in the standard normal table. Since the standard normal distribution is symmetric around the expectation 0, $P(|Z_n^*| \leq z) \approx P(-z \leq Z_n^* \leq z)$. Let Φ be the cumulative probability distribution of the standard normal distribution. For a confidence level of $\beta = 0.95$, for instance, we get $\Phi(1.96) \approx P(Z_n^* \leq 1.96) \approx 0.975$. Thus,

$$\begin{aligned} P(-1.96 \leq Z_n^* \leq 1.96) &\approx P(Z_n^* \leq 1.96) - P(Z_n^* \leq -1.96) \\ &\approx \Phi(1.96) - (1 - \Phi(1.96)) \\ &\approx 0.975 - 0.025 = 0.95. \end{aligned}$$

and $z \approx 1.96$ (see also the illustration in Fig. 2.5).

Confidence Interval. Since

$$\beta = P(|Z_n^*| \leq z) \approx P(|Z_n - \mu| \leq z\sqrt{S^2/n})$$

we get the *confidence interval*

$$\left[Z_n - z\sqrt{S^2/n}, Z_n + z\sqrt{S^2/n} \right] = [I_l, I_r].$$

We see now that the confidence interval depends on the number of generated trajectories n . To make the confidence interval twice as small, we must generate four times as many trajectories.

Note that we cannot deduce that μ lies in $[I_l, I_r]$ with probability β since

- Z_n and S^2 are random variables,
- Z_n is normally distributed only as $n \rightarrow \infty$,
- we approximated σ^2 by S^2 .

The correct interpretation is that, for large n and a large number of realizations of the random interval $[I_l, I_r]$, β is the percentage of realizations $[I_l(\omega), I_r(\omega)]$ that will contain μ .

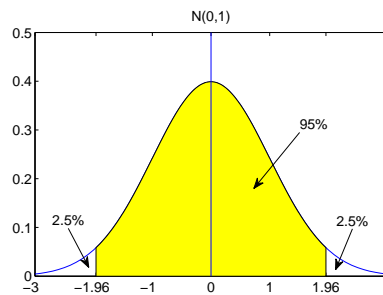


Figure 2.5: Probability density function of $\mathcal{N}(0, 1)$.

Practical Issues. The simulation algorithm in Section 2.2.4 can be used to compute realizations of the random variables X_i . Assume that we execute the algorithm n times to generate n trajectories. We add variables *sum* and

\widehat{sum} that we initialize with zero. In the i -th run, we increase sum and \widehat{sum} by $x_i = X_i(\omega)$ and $x_i^2 = (X_i(\omega))^2$, respectively. After the generation of n trajectories, we compute $z_n = sum/n$ and

$$s^2 = \frac{\widehat{sum}}{n-1} - \frac{sum^2}{(n-1)n}.$$

Note that in the latter case we exploit Eq. (2.3).

Then, for a fixed confidence, the interval

$$\left[z_n - z\sqrt{s^2/n}, z_n + z\sqrt{s^2/n} \right]$$

can be constructed.

Number of Simulation Runs. If the interval $[I_l, I_r]$ is large relative to Z_n the quality of the estimator is poor and more simulation runs have to be carried out. This is likely the case if n is small or if we try to estimate the probability of an event that is rare. Let us fix the relative width of the interval to be 0.2 (which means that we have a relative error of at most 0.1) and chose confidence level $\beta = 0.95$. Thus, $z \approx 1.96$. Assume that we want to estimate $\gamma = P(A)$ and X_i is one if A occurs in the i -th run and zero otherwise. Clearly,

$$E(X_1) = E(X_2) = \dots = E(X_n) = 1 \cdot \gamma + 0 \cdot (1 - \gamma) = \gamma.$$

We can determine the number of necessary runs by bounding the relative width of the confidence interval as follows:

$$2 \cdot \frac{z \cdot \sqrt{S^2/N}}{\gamma} \leq 0.2 \implies \frac{z^2}{0.01} \frac{S^2}{\gamma^2} \leq N \implies 384 \cdot \frac{S^2}{\gamma^2} \leq N$$

Using the fact that

$$\sigma^2 = \text{VAR}[\chi_A] = \gamma(1 - \gamma)$$

and replacing S^2 by σ^2 yields $N \geq 384 \cdot \frac{1-\gamma}{\gamma}$. For instance, the sufficient number of runs to guarantee that probabilities, having at least the order of magnitude of 10^{-5} , are estimated with a relative error of at most 0.1 and a confidence of 95% is $N \approx 38,000,000$.

2.4 Summary

The generation of trajectories is a simple method to get insight into the behavior of chemically reacting systems. Especially if the size of the model is large and the event whose probability has to be calculated occurs frequently, Gillespie simulation is preferred over numerical solution methods (e.g. ODE solvers). There are, however, some issues that require further attention:

- Gillespie simulation can be time consuming. Firstly, if the chemical reactions occur at time scales that differ by several orders of magnitude. In this case, many fast reactions have to be carried out until a slow reaction occurs. Secondly, if the event whose probability has to be calculated is rare, a large number of simulation runs is necessary to estimate its probability.
- In order to half the confidence interval, four times more simulation runs have to be carried out. It is therefore expensive to obtain accurate results.
- It is important that the pseudorandom numbers used during the simulation are close to "truly random". This is necessary to ensure independence. We refer to [LK00] for a detailed discussion.

Chapter 3

Numerical Solution Methods

3.1 Introduction

In this chapter we will consider a matrix-vector form of the CME in order to derive alternative expressions for the solution of the CME, that is, the state probabilities $P(X^{(t)} = x)$. We will see that, for finite systems, the solution of the CME equals the product of a vector and a matrix exponential. The calculation of the matrix exponential, however, is computationally expensive but it motivates an alternative solution method, called *uniformization*. For the solution of Markov chains, uniformization is often preferred over numerical integration methods which are the basis of most ODE solvers.

The numerical solution of the CME (by using an ODE solver or applying uniformization) gives us the probabilities of *all* states no matter how small this value is. This makes standard numerical solution methods infeasible if the state space of the models is extremely large (or even infinite). In this case, we can generate a number of trajectories using Gillespie simulation and estimate the probabilities of interest. Gillespie simulation provides statistical estimates, whereas a numerical solution usually provides an accurate approximation of the “true probabilities”. In the generated trajectories, we usually only see a small a part of the state space, namely those states that have a “sufficiently high” probability. This idea can also be used for numerical approximations of the solution. Similar to Gillespie simulation, we only explore the “significant parts” of the state space and compute approximations of the state probabilities $P(X^{(t)} = x)$ at a particular time instance $t \geq 0$.

3.2 Matrix-Vector Form and Matrix Exponential

The CME describes the change of the probability of a state x at time t as

$$\begin{aligned} \frac{d}{dt}P(X^{(t)} = x) &= \sum_{\substack{j=1, \\ x-v_j \geq 0}}^M \alpha_j(x-v_j) \cdot P(X^{(t)} = x-v_j) \\ &\quad - \sum_{j=1}^M \alpha_j(x) \cdot P(X^{(t)} = x) \\ &= \sum_{\substack{j=1, \\ x-v_j \geq 0}}^M \alpha_j(x-v_j) \cdot P(X^{(t)} = x-v_j) - \alpha_0(x) \cdot P(X^{(t)} = x). \end{aligned}$$

Let $f : \mathbb{N}_0^N \rightarrow \mathbb{N}_0$ be a bijective function that enumerates all states, that is, $x = f^{-1}(i)$ is the i -th state in the enumeration. We define the row vector

$$p^{(t)} = \left(P(X^{(t)} = f^{-1}(1)), P(X^{(t)} = f^{-1}(2)), \dots \right)$$

as a vector that contains all state probabilities at time t . Moreover, let $Q = (q_{ij})_{i,j \in \{1,2,\dots\}}$ be a matrix with entries

$$q_{ij} = \begin{cases} -\alpha_0(x) & \text{if } i = j, f(x) = i, \\ \alpha_m(x) & \text{if } f(x) = i, f(x+v_m) = j, \\ 0 & \text{otherwise.} \end{cases}$$

Then we can express the CME in terms of Q and $p^{(t)}$ by

$$\frac{d}{dt}p^{(t)} = p^{(t)} \cdot Q. \tag{3.1}$$

The matrix Q is called the *infinitesimal generator matrix* of X .

Each Markov chain has an associated state-transition graph, called *intensity graph*. It is a directed graph whose node set corresponds to the state space of the chain. It has an edge from state i to state j labeled by q_{ij} whenever $q_{ij} > 0$. The Markov chain is uniquely determined by its intensity graph (since the diagonal entry of Q is given by the negative sum of the edge labels of a state, i.e., $-\alpha_0(x) = -\sum_{m=1}^M \alpha_m(x)$).

The matrix-vector form of the CME in Eq. (3.1) describes the CME as a special case of a system of linear differential equations. For systems of linear differential equations of this form, however, the solution $p^{(t)}$ can be expressed not only as the unique vector that satisfies Eq. (3.1) (and some initial condition for $t = 0$), but also in an alternative way presented below. Let us fix the initial distribution

$$p^{(0)} = \left(P(X^{(0)} = f^{-1}(1)), P(X^{(0)} = f^{-1}(2)), \dots \right).$$

If Q is bounded in norm (e.g. $\sup_{ij} |q_{ij}| < \infty$) then the general solution of a linear system of differential equations of the form in Eq. (3.1) can be written as

$$p^{(t)} = p^{(0)} \cdot e^{Qt}.$$

If Q is of finite size, the matrix exponential e^{Qt} can be computed as

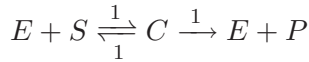
$$e^{Qt} = \sum_{i=0}^{\infty} \frac{(Qt)^i}{i!}.$$

Hence, if only finitely many states are reachable, that is, for each $t \geq 0$ the set $\{x \in \mathbb{N}_0^n \mid P(X^{(t)} = x) > 0\}$ is finite, we get

$$p^{(t)} = p^{(0)} \cdot e^{Qt} = p^{(0)} \cdot \sum_{i=0}^{\infty} \frac{(Qt)^i}{i!}. \tag{3.2}$$

Example 30: ENZYME KINETICS

Consider the CTMC $(X^{(t)})_{t \geq 0}$ for the enzyme kinetics reaction network



with initial state $X^{(0)} = (3, 2, 0, 0)$. Figure 3.1 shows the intensity graph of X where the states are numbered from 1 to 6. The infinitesimal generator matrix of X is

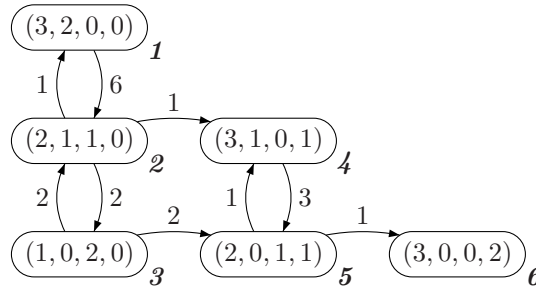


Figure 3.1: The intensity graph for the enzyme kinetics example.

$$Q = \begin{pmatrix} -6 & 6 & 0 & 0 & 0 & 0 \\ 1 & -4 & 1 & 2 & 0 & 0 \\ 0 & 2 & -4 & 0 & 2 & 0 \\ 0 & 0 & 0 & -3 & 3 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

3.3 Uniformization

For a given continuous-time Markov chain X , the uniformization of X is based on the construction of

- a discrete-time Markov chain Y that represents the sequence of states visited by X and
- a Poisson process N that represents the jump times of X .

In this section, we discuss these two concepts and combine them. We will abbreviate continuous-time Markov chain by CTMC and discrete-time Markov chain by DTMC in the following.

Poisson Processes. A CTMC $(N^{(t)})_{t \geq 0}$ is called a Poisson process with intensity λ if $N^{(t)} : \Omega \rightarrow \{0, 1, \dots\}$ and for an infinitesimal time interval $[t, t + \Delta)$,

$$P(N^{(t+\Delta)} = k + 1 \mid N^{(t)} = k) = \lambda \cdot \Delta$$

for all $t \geq 0$ and $k \in \{0, 1, \dots\}$. In this case the CME is

$$\frac{d}{dt}P(N^{(t)} = k) = \lambda \cdot P(N^{(t)} = k - 1) - \lambda \cdot P(N^{(t)} = k)$$

for $k \in \{1, \dots\}$ and for $k = 0$ we get

$$\frac{d}{dt}P(N^{(t)} = 0) = -\lambda \cdot P(N^{(t)} = 0).$$

Assume $P(N^{(0)} = 0) = 1$. Then

$$P(N^{(t)} = 0) = P(T_1 > t) = e^{-\lambda t}$$

where T_1 is the first jump time of N . By induction it can be shown that

$$P(N^{(t)} = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

Discrete-Time Markov Chains. Let S be a discrete set. A *discrete-time Markov chain* is a family of random variables $(Y_n)_{n \in \mathbb{N}_0}$, where n is a discrete index, $Y_n : \Omega \rightarrow S$ and Y fulfills the Markov property

$$P(Y_{n+1} = y \mid Y_n = x_n, Y_{n-1} = x_{n-1}, \dots, Y_0 = x_0) = P(Y_{n+1} = y \mid Y_n = x_n)$$

for all n, y, x_0, \dots, x_n .

We assume that the states in S are ordered by a function $f : S \rightarrow \mathbb{N}$ and that the transition probabilities $p_{ij} = P(Y_{n+1} = f^{-1}(j) \mid Y_n = f^{-1}(i))$ do not depend on n . Thus, we can arrange the transition probabilities in a matrix $P = (p_{ij})_{i,j \in \{1,2,\dots\}}$. The matrix P is called the transition probability matrix of DTMC Y .

Let p_0 be the row vector that contains the initial distribution, i.e., the entries $P(Y_0 = x)$ for all x . We now find that $p_1 = p_0 \cdot P$ since

$$P(Y_1 = x) = \sum_{y \in S} P(Y_1 = x \mid Y_0 = y) \cdot P(Y_0 = y).$$

Successive application of this argument yields

$$p_n = p_{n-1} \cdot P = \dots = p_0 \cdot P^n,$$

that is, the probabilities after n steps are obtained by n multiplications with P .

Note that P is a stochastic matrix, i.e., the row sums are one and it has only non-negative entries. Moreover, we can represent finite DTMCs by a graph in a similar way as CTMCs. In the discrete-time setting, self-loops are possible, namely, a state may have a positive transition probability to itself ($p_{ii} > 0$). In this case, we may have $Y_n(\omega) = Y_{n+1}(\omega)$. As opposed to that, for any CTMC X with jump times T_0, T_1, \dots we have $X^{(T_n)}(\omega) \neq X^{(T_{n+1})}(\omega)$ for all ω and all n .

Example 31: DTMC

Figure 3.2 shows the intensity graph of a DTMC Y . The matrix

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/3 & 0 & 2/3 \\ 1 & 0 & 0 \end{pmatrix}.$$

contains the transition probabilities of Y .

Let $p_0 = (1/3, 1/3, 1/3)$ be an initial distribution. Then, after two steps,

$$p_2 = p_0 \cdot P^2 \approx (0.5833, 0.3056, 0.1111),$$

that is, for instance, $P(Y_2 = 2) \approx 0.3056$.

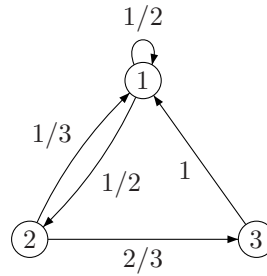


Figure 3.2: The intensity graph for the DTMC example.

Uniform Markov Chain. Let $(Y_n)_{n \in \mathbb{N}_0}$ be a DTMC with discrete state space S and let $(N^{(t)})_{t \geq 0}$ be a Poisson process with intensity $\lambda > 0$. Assume that N is independent of Y . For $t \geq 0$, we define the *uniform Markov chain* $(X^{(t)})_{t \geq 0}$ as

$$X^{(t)} = Y_{N^{(t)}}.$$

It is easy to see that X is a CTMC. The processes N is called the *clock* and Y is called the *subordinated DTMC*.

We compute

$$\begin{aligned} P(X^{(t)} = x) &= P(Y_{N^{(t)}} = x) \\ &= \sum_{n=0}^{\infty} P(Y_n = x, N^{(t)} = n) \\ &= \sum_{n=0}^{\infty} P(Y_n = x) \cdot P(N^{(t)} = n) \\ &= \sum_{n=0}^{\infty} P(Y_n = x) \cdot e^{-\lambda t} \frac{(\lambda t)^n}{n!}. \end{aligned}$$

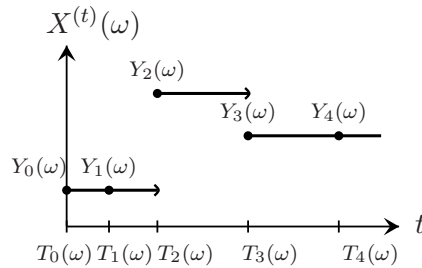


Figure 3.3: The jump times T_0, T_1, \dots of N and the states visited by Y yield X .

Let p_0 and P be the initial distribution and transition probability matrix of Y , respectively. In vector matrix form, we get

$$p^{(t)} = \sum_{n=0}^{\infty} p_0 \cdot P^n \cdot e^{-\lambda t} \frac{(\lambda t)^n}{n!} = \sum_{n=0}^{\infty} p_n \cdot e^{-\lambda t} \frac{(\lambda t)^n}{n!},$$

where $p^{(t)}$ contains the state probabilities of X and p_n the state probabilities of Y .

This means that the transition probabilities of a uniform Markov chain are given by the solution of the subordinated DTMC (vectors p_0, p_1, \dots) which are weighted with the Poisson probabilities and summed up.

Construction of the Clock and Subordinated DTMC. Our aim is now to construct for a given finite CTMC X a Poisson process N and a DTMC Y such that X can be expressed as the corresponding uniform Markov chain. Let Q be the infinitesimal generator matrix of X and assume that $\sup_{ij} |q_{ij}| < \infty$.

We choose $\lambda \geq \sup_{ij} |q_{ij}|$ and define $P = I + \frac{1}{\lambda}Q$, where I is the identity matrix. Then $Q = \lambda(P - I)$ and thus,

$$p^{(t)} = p^{(0)} \cdot e^{Qt} = p^{(0)} \cdot e^{\lambda t P - \lambda t I} = p^{(0)} \cdot e^{\lambda t P} \cdot e^{-\lambda t}.$$

Note that we exploited the fact that $P \cdot I = I \cdot P$ as well as the properties of the matrix exponential. By definition,

$$e^{\lambda t P} = \sum_{n=0}^{\infty} \frac{(\lambda t P)^n}{n!} = \sum_{n=0}^{\infty} P^n \frac{(\lambda t)^n}{n!}$$

and therefore

$$p^{(t)} = p^{(0)} \cdot e^{\lambda t P} \cdot e^{-\lambda t} = \sum_{n=0}^{\infty} p^{(0)} \cdot P^n \cdot e^{-\lambda t} \frac{(\lambda t)^n}{n!}. \quad (3.3)$$

It is important to note that $P = I + \frac{1}{\lambda}Q$ is a stochastic matrix, because for all $i, k \in \mathbb{N}$ with $i \neq k$,

$$\sum_{j \in \mathbb{N}} p_{ij} = 1 + \sum_{j \in \mathbb{N}} \frac{q_{ij}}{\lambda} = 1 + \frac{q_{ii}}{\lambda} + \sum_{j \in \mathbb{N}, j \neq i} \frac{q_{ij}}{\lambda} = 1.$$

and $p_{ik} = q_{ik}/\lambda \geq 0$ and $p_{ii} = 1 + q_{ii}/\lambda \geq 0$.

Here, we used the fact that the diagonal entry of a row in Q is the negative sum of the off-diagonal entries. Hence, P defines a DTMC Y and the vector

$$p_n = p^{(0)} \cdot P^n$$

contains the probabilities $P(Y_n = i)$ for all states i . Similarly, we can identify $e^{-\lambda t} \frac{(\lambda t)^n}{n!}$ as the probabilities of a Poisson process with intensity λ .

It remains the question why the infinite sum in Eq. (3.3) is more advantageous than the infinite sum in Eq. (3.2).

- The sum in Eq. (3.2) contains negative and nonnegative summands whereas in Eq. (3.3) all summands are nonnegative. Hence, in Eq. (3.3) round-off errors will occur less often.
- The infinite sum in Eq. (3.3) can be truncated w.r.t. an a priori specified accuracy $\epsilon > 0$. Consider a Poisson distributed random variable Z , that is, $P(Z = k) = e^{-\mu} \frac{\mu^k}{k!}$. Assume that $L, R \in \mathbb{N}_0$ are chosen such that, for $\mu = \lambda t$, $P(Z < L) = \epsilon/2$ and $P(Z > R) = \epsilon/2$. If we use L and R as truncation points, the approximation error for a state x becomes

$$\begin{aligned}
 & \sum_{n=0}^{\infty} P(Y_n = x) \cdot e^{-\lambda t} \frac{(\lambda t)^n}{n!} - \sum_{n=L}^R P(Y_n = x) \cdot e^{-\lambda t} \frac{(\lambda t)^n}{n!} \\
 = & \sum_{n < L, n > R} P(Y_n = x) \cdot e^{-\lambda t} \frac{(\lambda t)^n}{n!} \\
 \leq & \sum_{n < L, n > R} e^{-\lambda t} \frac{(\lambda t)^n}{n!} = \epsilon.
 \end{aligned}$$

- If the initial probability distribution is such that X starts in a certain state with probability one then the number of nonzero entries in p_1, p_2, \dots increase after each multiplication with P by a factor of at most M (number of reactions). In order to keep the number of nonzero entries in p_n small, we can set all entries to zero that are below a certain threshold $\delta > 0$. This introduces an additional approximation error, which turns out to be acceptably small in practice.
- The requirement $\lambda \geq \sup_{ij} |q_{ij}|$ can be relaxed since all states x with $P(Y_n = x) = 0$ for $n \in \{0, 1, \dots, R\}$ are not considered. Thus, it is sufficient to choose λ such that $\lambda \geq |q_{ii}|$ for states i within a “distance” of length R from the initial state¹.

3.4 Summary

In Markov models of realistic size, the number of states with significant probability can be large. Then Gillespie simulation as well as numerical approximation techniques are infeasible. This is especially the case if the system contains both species with small and species with large populations. The small populations are the reason why a stochastic model is necessary. On the other hand, the large populations in isolation can be easily analysed using reaction rate equations (provided that they are only marginally influenced by the small populations). If the evolution of small and large populations, however, depend on each other, a hybrid model is necessary to reduce the model complexity.

¹The term “distance” refers to the number of steps needed to reach state i with positive probability from the initial state.

3.4. SUMMARY

Further challenges related to stochastic models of biochemical reaction networks include

- sensitivity analysis methods to study the influence of model parameters such as reaction rate constants on the probability distributions,
- parameter estimation methods to estimate model parameters (such as reaction rate constants) if probability distributions at certain time instances are given.

Bibliography

- [ARM98] A. Arkin, J. Ross, and H. H. McAdams, *Stochastic kinetic analysis of developmental pathway bifurcation in phage λ -infected escherichia coli cells.*, Genetics **149** (1998), 1633–1648.
- [ELSS02] M. B. Elowitz, M. J. Levine, E. D. Siggia, and P. S. Swain, *Stochastic gene expression in a single cell*, Science **297** (2002), 1183–1186.
- [Gil92] D.T. Gillespie, *A rigorous derivation of the chemical master equation*, Physica A **188** (1992), 404–425.
- [HNW08] E. Hairer, S. Norsett, and G. Wanner, *Solving ordinary differential equations i: Nonstiff problems.*, Springer, 2008.
- [HW04] E. Hairer and G. Wanner, *Solving ordinary differential equations ii. stiff and differential-algebraic problems.*, Springer, 2004.
- [Kur72] T. G. Kurtz, *The relationship between stochastic and deterministic models for chemical reactions.*, J. Chem. Phys. **57** (1972), no. 7, 2976 –2978.
- [LK00] A. M. Law and W. D. Kelton, *Simulation modeling and analysis.*, McGraw Hill, 2000.