Information processing in stochastic two-component signal transduction

Master's thesis in Mathematics

Author:
D.G.A. Mondeel
1954113

Supervisors:
Dr. R. Planqué
Dr. J. Hulshof

Date:
July 5, 2014

Abstract

Cells face a fundamental problem: signals from the extracellular environment must be detected on the cellular surface and conveyed correctly to the interior for further processing. However, due to inherent stochasticity in the molecular reactions, this signal arrives imperfectly. This raises the question of how well a cell is able to track its changing environment and respond appropriately. This situation can be captured in a mathematical model of two random variables for the signal and internal output of signal transduction. Though several dependency or quality measures exist, it remains unclear how to best capture this ability mathematically. Mutual information is a popular, but we find that it has a key shortcoming in that it is not equipped to handle some interesting biological questions - for instance, the probability that the cell has a wrong internal representation. We show for a simplified gene expression model that mutual information and error probability do not always agree.

We will focus in particular on the two-component system signal transduction mechanism, using the linear noise approximation to investigate fluctuations in the output and calculating the mutual information between signal concentration and output concentration. Recently, a similar approach was used in the literature on a simplified model neglecting any protein complex formation. Here, we consider a model including complex formation and find significantly different results for output variance and mutual information dependence on the mean signal level.

Contents

Preface

From January to June 2014, I worked on this master's thesis on information processing in two-component signal transduction as the final stage of the master program in Mathematics at the VU University in Amsterdam. The aims of such a thesis project are to investigate a mathematical research problem culminating in the exploration of the boundaries of scientific knowledge, to study relevant papers from the literature, to combine these, and ideally to make an original contribution. This written work is the result of my attempt to achieve this.

During my studies, I focused on the interplay between mathematics and biology and cellular biology and systems biology in particular. I was thus motivated to choose a thesis subject that would originate from a relevant and interesting biological research question, but which was deeply intertwined with mathematical questions and techniques needed to answer it. Following Dr. Frank Bruggeman's suggestion, I came upon the subject of information processing in two-component signal transduction systems. This proved to be quite an intriguing subject centered around the very practical biological question: to what extent is a cell able to track its changing environment in the presence of noise? Attempting to answer this question in light of two-component signaling systems requires a broad range of mathematical techniques (from Fokker-Planck equations and information theory to statistical concepts) and raises questions such as: how do we model a two-component signal transduction system mathematically? Can we analytically calculate the "information" passed through a two-component system for a certain information measure? Which measure is best for the task at hand? etc. I hope to shed light on some of these questions in the rest of this thesis.

First and foremost, thanks go out to my main supervisor Dr. Robert Planqué for his critical insights during this project and all his help and guidance. Secondly, to Dr. Frank Bruggeman for suggesting this topic and supplying us with the necessary background material, literature and his own insights and guidance. Finally, I would like to thank the second reader Dr. Joost Hulshof. I hope this work will inspire some answers to the many questions still remaining in this field and some fruitful future collaborations.

1 Introduction

This chapter serves as an introduction to the rest of the text, culminating in the research questions that will be considered in the next chapters. First, we must introduce the biological backdrop in order to put the questions we will attempt to answer in context.

1.1 Scientific context

The central dogma in biology states that the genes on DNA are transcribed into mRNA, which is subsequently converted on ribosomes into amino acid sequences that form proteins. Thus we could say that the information flow is from DNA to proteins. But what about the other way around: does information also flow from proteins to DNA? The answer is of course yes, and this allows cells to adjust their gene expression levels based on internal and external conditions. This is precisely how multicellular organisms develop different types of cells that are specialized for specific functions even though they all share the same DNA blueprint.

The cellular processes responsible for the control of gene expression are referred to as gene regulatory processes or networks. This control may happen in a variety of ways, but for our purposes the most significant way is through transcription factors (TFs). Transcription factors are special proteins that are able to modify the expression of one or several genes by binding to the promotor site of a gene and allowing (or blocking) RNA polymerase to transcribe the DNA. The effect of the binding of a transcription factor can increase the gene expression rate (activation) or negatively influence the expression rate (repression). The latter may occur because the transcription factor is blocking the RNA polymerase and other molecules needed for transcription from binding and forming a complex.

Genetic regulatory networks allow cells to respond to varying internal and external conditions by changing gene expression levels over time. Before genes can be regulated however, the environmental signals must first be transduced to the intracellular environment. In this thesis, we will concern ourselves with a signal transduction mechanism referred to as a two-component system. We will consider the output of this system to be a transcription factor, which then goes on to influence the expression level of genes as described above. Through a two-component system, the concentration of a signal molecule in the extracellular fluid is correlated with the concentration of its output molecule. In a way, this output molecule represents the cell's internal knowledge of the concentration of the signal, which can be used to respond to the environment. Since molecular reactions and gene expression are noisy processes; the relationship between the signal and the internal output of signal transduction will be imperfect. This means that the cell constantly has an imperfect view of its environment. In this thesis we are interested in just how imperfect this view is. Put differently, how much information does a cell have about its environment and to what degree is it able to track changes in it? We will investigate the size

of fluctuations in this process and consider how they depend on the parameters of the system. More fundamentally we will ask: what do we mean by the term information in this context and can we quantify this term in a meaningful way? We will focus on these questions first in general, and later we will apply our realizations to a specific two-component signal transduction system.

Understanding these questions about the impact of stochastic fluctuations on the ability of cells to respond to the environment correctly is of fundamental importance to biology, as these fluctuations may propagate into the process of gene expression itself which plays a role in many other biological subjects, including disease.

As this is a Master's project in Mathematics many relevant mathematical details will be included. Although not strictly necessary to understand the results they do deepen our understanding of them.

1.2 Research questions

Summarizing, we will address the following research questions in this text:

- How can we quantify the knowledge or information a cell has about its environment? More specifically, how can we use information theory for this purpose?
- What is a good measure for the quality of this signal transduction process?
- How does mutual information relate to the probability of the cell's internal representation being wrong?
- What is the linear noise approximation and can we analytically and numerically calculate it for a specific two-component system?
- How does the system's output variance change with model parameters? Particularly, how does it change with the mean concentration of the extracellular signal?
- What consequences does this have for the mutual information between the input and output of the signal transduction process?
- How do our results compare to the recent publication [Maity et al., 2014]?

Of these, the first, second and fourth research questions are mainly literature studies.
In trying to answer these questions, we came across a whole range of mathematical techniques as well as a vast amount of literature. To make this thesis as self-contained as possible, the first two chapters will mostly deal with background material that does not directly relate to two-component systems. These will serve to introduce the modeling steps used later on. Specifically, we start out by considering the question of how to quantify the term "information". In doing so, information theory and mutual information will be introduced since it is the historically used and most popular definition of information. We will also consider possible alternatives. In Chapter 3 we will then focus on *stochastic* modeling of chemical events in the cell as opposed to deterministic modeling. Of particular importance for the following chapter is the linear noise approximation. In Chapter 4 we will consider two-component systems and attempt to apply the linear noise approximation to a specific model. We will in particular investigate the impact of adjusting model parameters on the output variance and compare this with recent literature. In Chapter 5 we explore decision making and gene expression based on the internal representation

of an extracellular signal. Finally, we will conclude with a summary and pointers for future work. Some details that would have steered the main text too far off course have been collected in the appendix.

2 | Information theory in biology

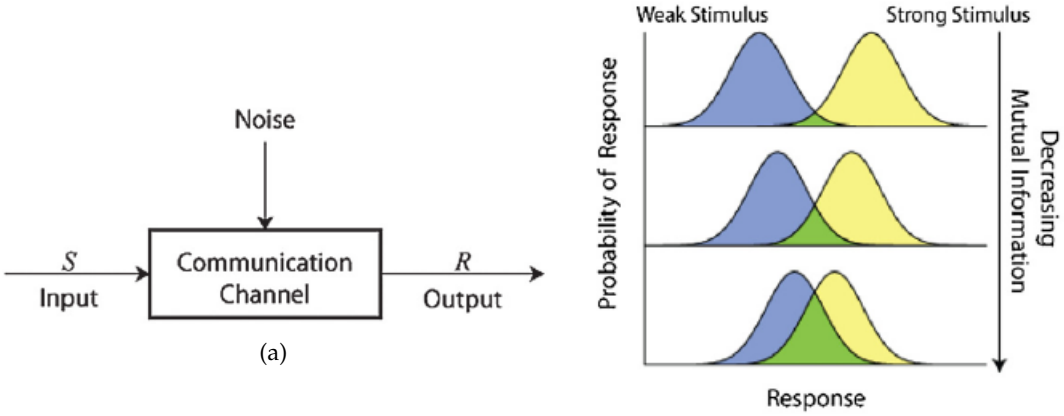
In this chapter we concern ourselves with the question of how can we model information transmission in biological signal transduction networks? Specifically, we ask: how do cells encode information about their environment? How can we model this, borrowing ideas from information theory? What measure can we use to judge the quality of this communication process?
We will turn to mutual information, which has its roots in information theoretic entropy and introduce all the necessary details as well as some recent developments.

2.1 Information and cellular signaling systems

One of the earliest papers in which information processing was considered in a biological context seems to be [Attneave, 1954] in the context of perceptual systems. The following is proposed: "A major function of the perceptual machinery is to strip away some of the redundancy of stimulation, to describe or encode information in a form more economical than that in which it impinges on the receptors". Another key paper was [Linsker, 1988], in a similar context, where the following idea was put forth: "The organizing principle I propose is that the network connections develop in such a way as to maximize the amount of information that is preserved when signals are transformed at each processing stage of the network, subject to certain constraints". Ever since, the consideration of information as a useful concept in biology has grown. Particularly in modeling of neurobiology it plays a large role [Rieke, 1999], but its applications are also growing in cell biology and genetics [Walczak and Tkačik, 2011].
Fundamentally, signal transduction in cells can be viewed as an information transmission problem, where chemical messengers relay information about the environment to the interior of a cell. After this signal arrives, some decision may have to be made about how to respond to new conditions. As pointed out in [Perkins and Swain, 2009], such cellular decision making must be probabilistic due to the three levels at which it occurs, all of which are, or at least can be, stochastic:

- Cells receive a noisy signal and must infer from that signal the state of the environment, now and in the future and
- They must weigh the costs and benefits of possible responses, given the probable future.
- They must make these decisions in the presence of other decision-makers, other cells.

Cells thus face a fundamental problem: signals from the environment must be detected on the surface of the cell and transduced to the internal decision-making parts. However, due to inherent stochasticity in the molecular reactions through which this occurs, this signal will arrive imperfectly. This problem is compounded by the fact that they must make decisions based on



(a) (b)

Figure 2.1: (a) **A communication channel**. A stochastic input signal (S) is sent through the channel with added noise resulting in a stochastic output (R). (b) **Stimulus overlap**. For sufficiently large noise, a cell cannot use its internal response to accurately discern which stimulus regime (weak or strong) was encountered. Therefore, noise results in a loss of information about the input. These figures are adapted from [Rhee et al., 2012].

this noisy knowledge. We can capture the essence of the signaling process in a so called communication channel, a simple visualization of which is displayed in Figure 2.1a. Such a communication channel can be modeled mathematically by two random variables that have some statistical dependence. Through this dependence measuring the output or response (R) results in gaining knowledge about the original signal (S). Noise hampers the quality of this communication and one way in which this can be of particular importance to cells is in their ability to discern various states of the enviroment that require different responses; see Figure 2.1b.

How should we quantify the strength of dependence or association between two random variables? Specifically, we would like to do so without bias for relationships of a specific form and say something about the quality of the signal transduction in the process. One way to think about this is that we would like our dependence measure to give equal value to relationships with an equal noise size. This notion, recently also referred to as "equitability" [Reshef et al., 2013], [Kinney and Atwal, 2014], does not yet have a definitive mathematical formalization.

The most obvious measure of statistical dependence between a signal S and a response R is the covariance $\mathrm{Cov}(S,R) = \mathbb{E}\left[(S-\bar{S})(R-\bar{R})\right]$, or the related Pearson correlation coefficient $\rho = \frac{\mathrm{Cov}(S;R)}{\mathrm{std}(S)\,\mathrm{std}(R)}$ (A.1). Unfortunately, ρ is not a useful measure of dependency in general. Firstly, statistical correlation does not guarantee a causal relationship, and vice versa, if there is such a relationship it does not guarantee existence of correlation. Secondly, it is best suited to continuous, normally distributed data and perhaps most importantly, it only detects linear relationships between random variables and will wrongly classify any non-linear relationship. A currently much-used alternative is mutual information, a fundamental quantity in information theory, which *can* detect non-linear relationships; see Figure 2.2.

As we expect the relationship between inputs and outputs in genetic regulatory networks to be non-linear, this is an important characteristic for us. Also, assumption-free measures can be applied in many more situations. Mutual information has its roots in information theory, which we will introduce in the next section.

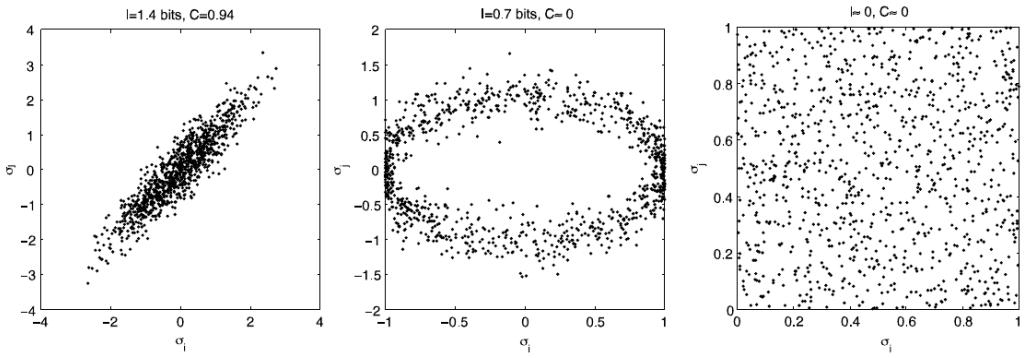


Figure 2.2: **Non-linear classification**. All three images are scatterplots of two variables drawn from three joint distributions. (Left) the variables are linearly related, and ρ is almost at its maximum 1. (Middle) The variables are dependent, butit is a non-linear dependecy. ρ takes the value 0, but mutual information gives a non-zero value showing its correct classification behaviour for non-linear relationships. (Right) The variables are independent, and both linear correlation and mutual information classify this as zero. Source: [Walczak and Tkačik, 2011].

2.2 Introduction to information theory

Information theory is a branch of applied mathematics started by C.E. Shannon in 1948. Initially developed for electrical engineers to design efficient and reliable communication systems, it has since spiralled out to become a field of research into the essence of the communication process itself and has found application in other fields such as neurobiology and more recently in genetic regulatory networks [Walczak and Tkačik, 2011]. In the next sections we introduce the main concepts of information theory and mutual information in particular.

2.2.1 Derivation of the Shannon entropy

The fundamental concept in information theory is that of *entropy*, which bears resemblance to, but is not the same as, the statistical physics concept. Consider a discrete random variable X, of which each realization may be a signal, and whose value may be uniformly quantized into a finite number of levels

$$X = \{x_k | k = 0, \pm 1, \pm 2, \ldots, \pm K\}.$$

Of course, if we let the spacing δx between the values go to zero and let K grow to infinity X becomes a continuous random variable and the sums in the definitions below become integrals (with some subtleties we will consider later). For the probability mass function of X, we use the notation $p_X(x_k) = \mathbb{P}(X = x_k)$. We distinguish this in notation from the probability density function of a continuous random variable X: $f_X(x)$.
There are two ways to motivate the definition of the Shannon entropy: by assuming the characteristics we want it to have and rigorously proving what form entropy thus has to take, or by intuitive reasoning showcasing the usability of the entropy. We will discuss both.

Rigorous derivation

Three assumptions underlie the derivation of the Shannon entropy as derived in Shannon's original paper [Shannon, 1948].

Theorem 2.2.1. Suppose we have a set of possible events whose probabilities of occurrence are $p_1, \ldots, p_N$ and that this is all we know about which event will occur. We are looking for a measure H for how uncertain we are of the outcome that satisfies the following 3 criteria:

1. **[Continuity]** H must be a continuous function of the probabilities of the outcome
2. **[Monotonicity]** When outcomes are equally likely, H should be a monotonically increasing function of N, the number of outcomes.
3. **[Additivity]** Multi-part choices may be visualized as branching trees and H should be the sum of the information gained at each branch point.

The only measure H that satisfies the assumptions mentioned above is the entropy of a random variable X with a probability mass function $p_X(x)$ defined by

$$H(X) = -\sum_{x_k} p_X(x_k) \log p_X(x_k). \tag{2.1}$$

Or, when X is a continuous random variable with probability density function $f_X(x)$ and support S

$$h(X) = -\int_S f_X(x) \log f_X(x) dx. \tag{2.2}$$

Proof. See Appendix 2 in [Shannon, 1948]. The proof is a little too long and non-insightful for our purposes to be reproduced here. □

Note that we distinguish in notation between the discrete and continuous cases. In the literature the continuous case is often referred to as differential entropy and denoted with h instead of H. Also, a little care has to be taken with summing over all x values in the definition since some values may have a probability mass of zero. In such cases we invoke the calculus result $\lim_{x \to 0^+} x \ln x = 0$, i.e. terms with zero probability do not influence the entropy.
For more than two variables the concept of entropy is straightforwardly extended; however, we will not be needing these extensions so they are not included.

Intuitive derivation

The concept of entropy is intricately related to the notion of surprising outcomes of an experiment. Suppose $p_k = 1$, and therefore $p_i = 0$ for $i \neq k$. In this case there is no surprise at the outcome of an experiment and no information is gained since we know what the signal will be. However, when p_k is small, there is more uncertainty and more surprise when X takes the value x_k. Let us define the amount of surprise after observing $X = x_k$ by

$$\mathrm{S}(x_k) = \log\left(\frac{1}{p_k}\right) = -\log p_k.$$

Now note that the surprise is again a discrete random variable which has value $\mathrm{S}(x_k)$ with probability p_k. Therefore, the mean of this random variable is

$$\begin{aligned} \mathbb{E}\,\mathrm{S}(x_k) &= \sum_k p_k \,\mathrm{S}(x_k) && (2.3)\\ &= -\sum_k p_k \log p_k && (2.4)\\ &= H(X). && (2.5) \end{aligned}$$

One can thus interpret entropy as the average amount of surprise per realization of X.

Theorem 2.2.2. The discrete version of the Shannon entropy satisfies the following three properties (a) Entropy is non-negative: $H(X) \geq 0$.
(b) Entropy is only zero when a single outcome has probability 1.
(c) $H(X)$ can be converted from one base to another by multiplying by an appropriate factor: $H_b(X) = \log_b(a) H_a(X)$.

Proof. (a) Note that $0 \leq p_X(x) \leq 1$ so that $-p_X(x)\log(p_X(x)) \geq 0$. Thus $H(X) \geq 0$.
(b) All zero probability terms add zero to the sum because of the $p_X(x) = 0$ term. For the remaining term the logarithm is zero.
(c) This is a trivial consequence of the logarithmic rule $\log_b(p_X(x)) = \log_b(a)\log_a(p_X(x))$. □

The base of the logarithm chosen determines the units involved, two of which are used especially often and prove useful: logarithms to base 2 resulting in units called *bits* and the natural logarithm resulting in so-called *nats*.
To see where the view of entropy in terms of an expectation comes in handy, consider the next example which we will come to use later on.

Example 2.2.1 (Entropy of a Gaussian distribution)**.** Assume some random variable X has a Gaussian distribution with density $\phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. Then the entropy (using the natural logarithm) is given by:

$$\begin{aligned} h(\phi(x)) &= -\mathbb{E}\ln\phi(x) = -\mathbb{E}\left[-\frac{(x-\mu)^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right] \\ &= \frac{\mathbb{E}(x-\mu)^2}{2\sigma^2} + \frac{1}{2}\ln(2\pi\sigma^2) \\ &= \frac{1}{2}\ln(e) + \frac{1}{2}\ln(2\pi\sigma^2) \\ &= \frac{1}{2}\ln(2\pi e\sigma^2) \text{ nats}, \end{aligned} \tag{2.6}$$

where we used the definition of variance $\mathrm{Var}(X) = \mathbb{E}(X-\mu)^2$. There are some remarks to be made here. First of all, this equation shows that the entropy is independent of the mean. This makes sense, since entropy should measure uncertainty and the mean of the distribution simply corresponds to an arbitrary choice of the origin of our coordinate system. Note that the entropy does depend on the variance σ^2 which makes sense for the same reason. In fact, if we double the standard deviation of X so that $\sigma \to 2\sigma$ then we see

$$\begin{aligned} h(X^*) &= \frac{1}{2}\frac{1}{\log_2(e)}\log\left(4(2\pi e\sigma^2)\right) \\ &= \frac{1}{2}\frac{1}{\log_2(e)}\left(2\log 2 + \log(2\pi e\sigma^2)\right) \\ &= \frac{1}{\log_2(e)} + h(X), \end{aligned}$$

meaning the entropy rises by one bit divided by the scaling factor. We see that entropy differences thus measure by which factor the variance has scaled.

2.2.2 The relationship between differential and Shannon entropy

Passing from the Shannon to differential entropy is actually quite subtle. First, begin by considering X to take the values $x_k = k\delta x$ and let $\delta x \to 0$. Also note that the continuous random variable takes on a value in $[x_k, x_k + \delta x]$ with probability $f_X(x_k)\delta x$. In the limit we can therefore write the discrete entropy of the continuous version of X as

$$\begin{aligned} H(X) &= -\lim_{\delta x \to 0} \sum_{-\infty}^{\infty} f_X(x_k)\delta x \log(f_X(x_k)\delta x) \\ &= -\lim_{\delta x \to 0} \left[\sum_{-\infty}^{\infty} f_X(x_k)\delta x \log(f_X(x_k)) + \sum_{-\infty}^{\infty} f_X(x_k)\delta x \log(\delta x) \right] \\ &= -\int_{-\infty}^{\infty} f_X(x) \log f_X(x) dx - \lim_{\delta x \to 0} \log \delta x \int_{-\infty}^{\infty} f_X(x) dx \\ &= h(X) - \lim_{\delta x \to 0} \log \delta x \end{aligned} \tag{2.7}$$

As $\lim_{\delta x \to 0} \log \delta x$ goes to infinity we see that a continuous random variable has infinite discrete entropy. This means that differential entropy is not the limit of the discrete entropy for $n \to \infty$ but differs from it by an infinite offset. The idea here is that this infinite term serves as a reference. In practice, the (mutual) information through a stochastic system is a difference between two entropy terms that now have a common reference and cancel out.
Differential entropy is in even more ways quite a troublesome concept since it differs from the ordinary Shannon entropy in more aspects, as shown by the following example.

Example 2.2.2. Consider a random variable X with a uniform distribution on $[a, b]$

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

Then

$$h(X) = -\int_a^b \frac{1}{b-a} \log\left(\frac{1}{b-a}\right) dx = \log(b-a).$$

This example shows that differential entropy can be negative for $(b-a) < 1$ whereas discrete entropy is non-negative (2.2.2). Secondly, it shows that for $b - a = 1$ differential entropy is zero.

Theorem 2.2.3. Differential entropy satisfies the following properties:
(a) It is translation invariant: $h(X \pm c) = h(X)$
(b) It is not invariant under scaling: $h(aX) = h(X) + \log(|a|)$.

Proof. (a) The definition of differential entropy does not involve the actual values of x. Only the probabilities. (b) This is proved by changing variables to $Y = aX$. A probability density function $f_X(x)$ integrates to one, so that $f_Y(y) = \frac{1}{|a|} f_X(y/a)$. But then we have that

$$\begin{aligned} h(Y) &= -\mathbb{E}\left[\log f_Y(y)\right] \\ &= -\mathbb{E}\left[\log f_X(y/a)\right] + \log(|a|) \\ &= h(X) + \log(|a|). \end{aligned}$$

□

2.2.3 Joint and conditional entropy

The concepts in this section will be defined for the discrete version of entropy. For the continuous case simply replace sums by integrals and probability mass functions by probability density functions. The definition for the joint entropy of a pair of random variables is quite similar to the one variable case (2.2.1). This makes sense when viewing the pair (X, Y) as a single vector-valued random variable.

Definition 2.2.1. The joint entropy of a pair of random variables (X, Y) with joint probability mass function $p_{X,Y}(x, y)$ is defined as

$$H(X, Y) = -\sum_{x,y} p_{X,Y}(x, y) \log p_{X,Y}(x, y) = -\mathbb{E} \log p_{X,Y}(x, y) \tag{2.8}$$

Now having definitions for marginal and joint entropy, we can also consider conditional entropy.

Definition 2.2.2. The conditional entropy of the random variable X given Y is defined as

$$H(X|Y) = -\sum_{x,y} p_{X,Y}(x, y) \log p_{X|Y}(x|y) = -\mathbb{E} \log p_{X|Y}(x|y) \tag{2.9}$$

Using the definitions above we are now ready to prove the following useful theorem, which is sometimes referred to as the chain rule.

Theorem 2.2.4. The joint, conditional and marginal entropy of a pair of random variables (X, Y) are related through

$$H(X, Y) = H(X) + H(Y|X) \tag{2.10}$$
$$= H(Y) + H(X|Y). \tag{2.11}$$

Proof. We prove the first equality; the second follows through symmetry.

$$\begin{aligned} H(X, Y) &= -\mathbb{E} \log p_{X,Y}(x, y) \\ &= -\mathbb{E} \left[\log \left(p_{Y|X}(y|x) p_X(x) \right) \right] \\ &= -\mathbb{E} \left[\log p_{Y|X}(y|x) + \log p_X(x) \right] \\ &= -\mathbb{E} \log p_{Y|X}(y|x) - \mathbb{E} \log p_X(x) \\ &= H(X) + H(Y|X) \end{aligned}$$

□

Figure 2.3 displays a nice graphical summary of the relationships between the various forms of entropy derived in the preceding theorems. The graph might seem to suggest a similarity between the entropy H and basic set theory. In fact, this similarity is not mere happenstance (see [Yeung, 2010] Chapter 3 for details). Figure 2.3 also includes the mutual information $I(X; Y)$ which is the subject of the next section.

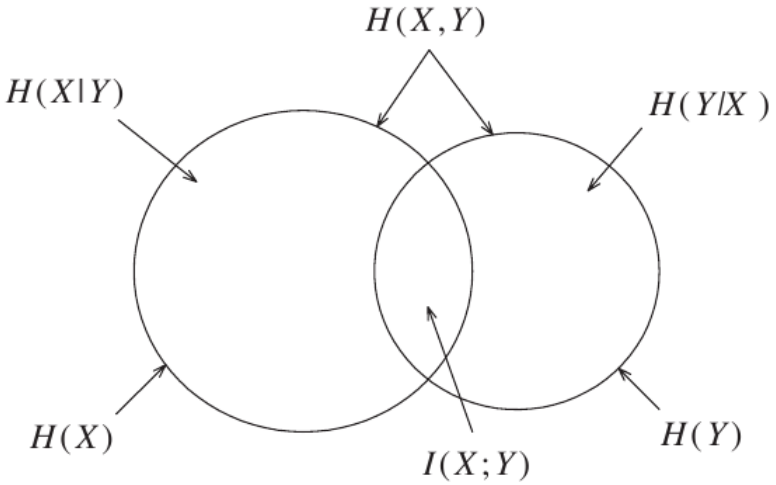


Figure 2.3: **An overview of the overlap between forms of entropy**. The two circles represent the marginal entropy of X and Y. Their union is $H(X,Y)$ and using theorem 2.2.4 we represent the partial circles as $H(Y|X)$ and $H(X|Y)$. The intersection is the mutual information $I(X;Y)$ Source: [Yeung, 2010].

2.3 Mutual information

Mutual information quantifies the concept of how much information one random variable (the response Y) contains about another random variable (the signal X). The mutual information for random variables X and Y is defined as (also see Figure 2.3)

$$I(X;Y) = H(X) - H(X|Y). \tag{2.12}$$

Or, in the continuous case

$$I(X;Y) = h(X) - h(X|Y).$$

Actually, from Figure 2.3 we can deduce several other equivalent definitions of $I(X;Y)$:

$$\begin{aligned} I(X;Y) &= H(X,Y) - H(X|Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y). \end{aligned}$$

What does this definition (2.12) mean? Remember that $H(X)$ measures our uncertainty about the random variable X. Similarly, $H(X|Y)$ measures the remaining uncertainty when we know the value of Y. Thus, $I(X;Y)$ is the reduction in uncertainty about X after observing the value of Y. This interpretation makes it a suitable measure for the quality of signal transduction, since it says how much the cell's uncertainty about the environment is reduced by knowing an internal concentration.

The reason for denoting the mutual information as $I(X;Y)$ and not as $I(X|Y)$ or $I(Y|X)$ is that

it is a symmetric quantity. To see this, consider the following:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= -\sum_x p_X(x)\log p_X(x) + \sum_{x,y} p_{X,Y}(x,y)\log p_{X|Y}(x|y) \\ &= -\sum_{x,y} p_{X,Y}(x,y)\log p_X(x) + \sum_{x,y} p_{X,Y}(x,y)\log p_{X|Y}(x|y) \\ &= \sum_{x,y} p_{X,Y}(x,y)\log \frac{p_{X|Y}(x|y)}{p_X(x)} \\ &= \sum_{x,y} p_{X,Y}(x,y)\log \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}. \end{aligned} \tag{2.13}$$

Notice that the last equality shows that the mutual information is symmetric with respect to Y and X. Observe that the symmetry of $I(\cdot)$ could have been deduced from Figure 2.3 too. This symmetry property has the nice consequence that we have two different but equivalent definitions of mutual information

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X). \tag{2.14}$$

This is very useful since it is often easier in biology to measure the response of the system to a variety of simulated signal regimes rather than measuring the actual signal. Also note that $I(X;Y)$ is completely determined by the joint distribution of the two random variables.
Previously we showed that H could be written in terms of the expectation operator; we can do the same for the mutual information. Considering (2.13) we see that

$$I(X;Y) = \mathbb{E}\log \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}. \tag{2.15}$$

What can we say about the range of values the mutual information can take? We saw in (2.2.2) that H (as opposed to h) cannot be negative. In addition, note that in general $H(X) \geq H(X|Y)$, so that $I(X;Y) \geq 0$. This lower bound is attained when X and Y are statistically independent as can be seen from (2.15). This might happen when, for instance, the noise is very large and thus the input and output are statistically independent.
Considering the above, we also see that $H(X|Y) \geq 0$ and thus we have that $I(X;Y) \leq H(X)$. However, due to symmetry (2.14) we also have $I(X;Y) \leq H(Y)$.
Summarizing,

$$0 \leq I(X;Y) \leq \min\{H(X), H(Y)\} \tag{2.16}$$

The upper bound is attained in a noiseless channel such that $H(X|Y) = 0$. This last result is particularly revealing since (as pointed out in [Rhee et al., 2012]) consequently the range of values that X and Y can take limit the communication capacity of the channel. If $H(Y) = 2$ bits then even though $H(X)$ maybe as high as 10 bits, $I(X;Y) \leq 2$ bits.

2.3.1 Kullback-Leibler divergence

There exists a generalization of mutual information called the relative entropy or Kullback-Leibler (KL) divergence. KL divergence of a distribution g from a distribution f for a random variable X, denoted $D_{KL}(f||g)$, is a measure of the information lost when g is used to approximate f. Think of g as the distribution we fit to the model while f is the true distribution. Specifically,

Definition 2.3.1. In the discrete case D_{KL} is defined by

$$D_{KL}(f||g) = \sum_{x_k} f(x_k) \ln \frac{f(x_k)}{g(x_k)}, \tag{2.17}$$

and in the continuous case when f has support S

$$D_{KL}(f||g) = \int_S f(x) \ln \frac{f(x)}{g(x)} dx. \tag{2.18}$$

This quantity is only defined if f and g are both normalized and $g(x) = 0$ implies $f(x) = 0$ for all x. Again, as for entropy, the convention $0 \ln 0 = 0$ is adopted because $\lim_{x \to 0^+} x \ln x = 0$.
The KL divergence is non-negative, this is sometimes referred to as the Gibb's inequality. Another fundamental result for KL divergence is that it is transformation invariant, as opposed to the differential entropy 2.2.3. For brevity we exclude the proofs, but they are easily found in online sources.
Note that mutual information (2.13) is a special case of KL divergence:

$$I(X;Y) = D_{KL}\left(p_{X,Y}(x,y)||p_X(x)p_Y(y)\right). \tag{2.19}$$

Another way to write this relationship is

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p_{X,Y}(x,y) \ln \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} \\ &= \sum_y p_Y(y) \sum_x p_{X|Y}(x|y) \ln \frac{p_{X|Y}(x|y)}{p_X(x)} \\ &= \sum_y p_Y(y) D_{KL}\left(p_{X|Y}(x|y)||p_X(x)\right) \\ &= \mathbb{E}_Y D_{KL}\left(p_{X|Y}(x|y)||p_X(x)\right). \end{aligned} \tag{2.20}$$

In other words, the mutual information $I(X;Y)$ is the expectation of the KL divergence of the marginal distribution $p_X(x)$ of X from the conditional distribution $p_{X|Y}(x|y)$ of X given Y. This illustrates the property that the more different the distributions $p_{X|Y}(x|y)$ and $p_X(x)$ are, the greater the increase in information.
Because of its non-negativity and invariance properties, KL divergence is seen as an improved generalization of differential entropy. From these properties one can deduce once again that mutual information is non-negative and transformation invariant.

2.3.2 The data processing inequality

So far we have seen that mutual information is a symmetric measure and that $0 \leq I(X;Y) \leq \min\{H(X), H(Y)\}$; let us now investigate other more subtle properties. Mutual information satisfies another fundamental result in information theory: the data processing inequality, which we will discuss next.

Definition 2.3.2. The random variables X, Y and Z form a Markov chain $X \to Y \to Z$ if the joint probability mass function can be written as

$$p_{X,Y,Z}(x,y,z) = p_X(x)p_{Y|X}(y|x)p_{Z|Y}(z|y). \tag{2.21}$$

Observe that this is a simple consequence of using the probabilistic chain rule $\mathrm{P}\left(\bigcap_{k=1}^{n} A_k\right) = \prod_{k=1}^{n} \mathrm{P}\left(A_k \middle| \bigcap_{j=1}^{k-1} A_j\right)$ and using the Markov property $X \perp Z|Y$. In addition, note that if $X \to Y \to Z$ forms a Markov chain then $Z \to Y \to X$ does too and we can write $X \leftrightarrow Y \leftrightarrow Z$.

Theorem 2.3.1 (Data processing inequality (DPI))**.** A dependence measure $D[X;Y]$ satisfies DPI if and only if whenever the random variables X, Y and Z form a Markov chain $X \leftrightarrow Y \leftrightarrow Z$

$$D[X;Y] \geq D[X;Z]. \tag{2.22}$$

Mutual information satisfies the DPI.

Proof. To prove this we actually need the chain rule for mutual information which we will use without proof [Cover and Thomas, 1991]

$$I(X_1; X_2; \ldots; X_N; Y) = \sum_{i=1}^{N} I(X_i; Y|X_{i-1}, X_{i-2}, \ldots, X_1).$$

Using the chain rule and the symmetry of $I(\cdot)$ we can expand the mutual information between X, Y and Z in 6 ways by reordering the elements. Take the following two expansions

$$\begin{aligned} I(X;Y;Z) &= I(X;Z) + I(X;Y|Z) \qquad \text{Order: Z,Y,X} \\ &= I(X;Y) + I(X;Z|Y) \qquad \text{Order: Y,Z,X.} \end{aligned} \tag{2.23}$$

By assumption Z is independent of X given Y, so that $I(X;Z|Y) = 0$. By virtue of the non-negativity of mutual information we have $I(X;Y|Z) \geq 0$, so that $I(X;Y) - I(X;Z) \geq 0$. We conclude

$$I(X;Y) \geq I(X;Z)$$

□

This means that manipulation of Y, cannot ever increase the amount of information that Y contains about X.

2.4 Optimizing mutual information

An especially nice property of mutual information is that it can be determined without a teacher making it ideal for self-organizing systems like neurons and other cells. Following the lines of reasoning used in neural network research [Haykin, 1994], we may want to set the mutual information as the objective function to be optimized, the so-called infomax principle due to Linsker [Linsker, 1988]. This information maximization is formalized in information theory by the concept of *channel capacity*.

2.4.1 Channel capacity

Choosing a communication channel in effect means choosing a distribution $p_{Y|X}(y|x)$. Since the mutual information (2.13) is fully specified by $p_{X,Y}(x,y) = p_{Y|X}(y|x)p_X(x)$, we are left with the ability to vary our choice for $p_X(x)$ which is not a property of the channel but of the signal. The channel capacity is defined as

$$C(X;Y) = \sup_{p_X(x)} I(X;Y). \tag{2.24}$$

The channel capacity is thus the maximal mutual information under all possible signal distributions $p_X(x)$. Biologically this quantity is relatively easy to ascertain. By supplying the system with a range of stimuli and sampling the response for each stimulus, we easily find the distribution $p_{Y|X}(y|x)$. The trick is in knowing which $p_X(x)$ are biologically realistic - this may or may not be known. Through the channel capacity we can give an upper bound on the information transfer through the channel.
The channel capacity actually has a very nice interpretation as the logarithm of the number of distinguishable input states, see the argument in [Cover and Thomas, 1991] Section 7.4. This means that using the formula

$$\text{No. distinguishable states} = 2^{I(X;Y)}$$

we can view mutual information in this sense too, a fact we will use later on since it allows us to say something about the number of environmental regimes that can be distinguished through signal transduction.
In general, determining the mutual information between an input X and output Y is a very difficult task. However, under special circumstances some progress can be made.

2.4.2 Mutual information in a Gaussian channel

In the special case of a so called Gaussian channel, a simple formula can be deduced for the mutual information through the channel. We start by assuming that the input/output relationship between X and Y is linear with additive Gaussian noise, meaning

$$Y = gX + Z,$$

where g is called the gain and Z is white noise (independent of X), i.e. has a Gaussian distribution with mean equal to zero and a variance σ_z^2:

$$f_Z(z) = f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi\sigma_z^2}} \exp\left(-\frac{z^2}{2\sigma_z^2}\right) \tag{2.25}$$

$$= \frac{1}{\sqrt{2\pi\sigma_z^2}} \exp\left(-\frac{(y-gx)^2}{2\sigma_z^2}\right). \tag{2.26}$$

It is quite reasonable for many biological situations to assume that the noise is Gaussian. Gaussian noise can arise from a fundamental physical reason or because it arises from the central limit theorem when many independent sources of noise are involved. However, the main reason for this assumption is of course its mathematical tractability.
If we now assume that the signal X also has a Gaussian distribution (with mean $\langle x \rangle$) then Y too must have a Gaussian distribution with mean $\langle y \rangle = g\langle x \rangle$ and variance $\sigma_Y^2 = g^2\sigma_X^2 + \sigma_z^2$. Under this *Gaussian channel assumption* we have, starting from the continuous version of (2.13):

$$\begin{aligned} I(X;Y) &= \int \mathrm{dx} \int \mathrm{dy}\, f_{X,Y}(x,y) \log_2 \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} \\ &= \frac{1}{\ln 2} \int dx \int dy f_{X,Y}(x,y) \ln \frac{f_{Y|X}(y|x)}{f_Y(y)} \\ &= \frac{1}{\ln 2} \mathbb{E} \ln \frac{f_{Y|X}(y|x)}{f_Y(y)}. \end{aligned}$$

Plugging in the distributions we find

$$I(X;Y) = \frac{1}{\ln 2}\langle \ln \left[\frac{\sqrt{2\pi\sigma_y^2}}{\sqrt{2\pi\sigma_z^2}}\right] - \frac{z^2}{2\sigma_z^2} + \frac{(y-\langle y\rangle)^2}{2\sigma_y^2}\rangle$$
$$= \frac{1}{\ln 2}\left[\ln\sqrt{\frac{\sigma_y^2}{\sigma_z^2}} - \frac{\langle z^2\rangle}{2\sigma_z^2} + \frac{\langle (y-\langle y\rangle)^2\rangle}{2\sigma_y^2}\right].$$

Since Z has mean zero and using the variance decomposition $\sigma_z^2 = \langle z^2\rangle - \langle z\rangle^2$, we see that $\langle z^2\rangle = \sigma_z^2$ so the second term simplifies to $-\frac{1}{2}$. In the third term the numerator is per definition the variance of Y so that it too simplifies to $\frac{1}{2}$. We conclude

$$\begin{aligned} I(Y;X) &= \frac{1}{\ln 2}\ln\sqrt{\frac{\sigma_y^2}{\sigma_z^2}} \\ &= \frac{1}{\ln 2\log_2(e)}\frac{1}{2}\log_2\frac{\sigma_y^2}{\sigma_z^2} \\ &= \frac{1}{2}\log_2\left(1+\frac{g^2\sigma_X^2}{\sigma_z^2}\right). \end{aligned} \tag{2.27}$$

We can rewrite (2.27) in a particularly nice way by considering the system as adding noise to the input directly:

$$Y = g(X+\eta),$$

where we could interpret η as the "effective" noise Z/g. Rewriting (2.27) we find

$$\begin{aligned} I(Y;X) &= \frac{1}{2}\log\left(1+\frac{\sigma_X^2}{\sigma_\eta^2}\right) \\ &= \frac{1}{2}\log\left(1+\text{SNR}\right), \end{aligned} \tag{2.28}$$

where $\frac{\sigma_X^2}{\sigma_\eta^2}$ is called the signal-to-noise ratio (SNR). Note that (2.28) captures the intuitive result that more noise should mean less information, as in Figure 2.1. We could of course also have directly written (2.27) in terms of the signal-to-noise ratio $SNR = \frac{\sigma_X^2}{\sigma_z^2/g^2}$.
In the appendix we derive another representation of the mutual information in terms of the correlation coefficient ρ, $I(Y;X) = -\frac{1}{2}\log\left(1-\rho^2\right)$; see the derivation of equation (A.3).

Derivation through entropy

An easier but less illuminating route to formula (2.28) exists by looking directly at the entropy of a normal distribution which is given by $\frac{1}{2}\ln(2\pi e\sigma^2)$; see example 2.2.1. Starting with (2.14) and realizing that X and η are normally distributed, we see

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= \frac{1}{2}\ln(2\pi e\sigma_Y^2) - \frac{1}{2}\ln(2\pi e\sigma_\eta^2) \\ &= \ln\frac{\sigma_Y^2}{\sigma_\eta^2} = \frac{1}{2}\ln\left(1+\text{SNR}\right) \end{aligned}$$

What have we learned? Maximization of information in a Gaussian channel can now be seen to be equivalent to maximizing the signal-to-noise ratio, which implies that we must either maximize the variability in the input for fixed noise variance or minimize the variance of the noise. In addition, maximization of information is also equivalent to maximizing the correlation coefficient, i.e. $\rho \to \pm 1$ see (A.1.1). Both of these results are intuitively satisfying.

2.4.3 Example: optimizing MI in Drosophilia development

Drosophila, a genus of flies commonly referred to as fruit flies, has been investigated in relation to mutual information in its embryonic development [Tkacik et al., 2008]. Upon production of a Drosophila egg, the mother stores Bicoid mRNA in the anterior portion of the egg which after translation into protein diffuses to the posterior side. This results in a Bicoid gradient along the anterior-posterior axis of the egg.
Bicoid functions as a morphogen, a signaling molecule that acts on cells and produces a response dependent on the local morphogen concentration. In the early stages of development the egg is filled with undifferentiated cells. The bicoid gradient along the long axis of the egg results in differential gene expression in a set of genes referred to as gap genes. This differential expression results down the line in spatial positioning determining the kind of a cell the undifferentiated cells develop to be.
By experimentally measuring the concentration of Bicoid c and the expression levels of the gap genes g simultaneously, one has the empirical joint probability distribution $p(c, g)$. With this and (2.13) one can calculate the mutual information between g and c. This turns out to be $I(c; g) = 1.5 \pm 0.1$ bits. It was traditionally thought that the gap genes have a switch-like response to bicoid. However, the 1.5 bits suggests more information than a 1 bit on-off response. To put this number into perspective, in [Walczak and Tkačik, 2011] the maximal mutual information (channel capacity) through this channel is considered. Numerically optimizing the Langrangian

$$\mathcal{L}[p(c)] = I(c; g) - \lambda \int p(c)dc \tag{2.29}$$

they find a maximum mutual information of 1.7 bits indicating that the experimentally measured information transmission is remarkably high! They use this finding in their investigation of using the maximization information as a possible design principle for genetic networks.

2.5 Recent developments concerning mutual information

Mutual information is not a perfect dependence measure. In fact, there are three difficulties stopping its full acceptance. The first is that it is rather difficult to estimate mutual information correctly from a small amount of data. This plays more of a role in data analysis than in what we are considering. Second, unlike ρ, mutual information does not come with an automatic interpretation of its values. A value of $\rho = 0.5$ says something about the spread of data points, $I(X; Y) = 2.2$ does not. Again, this is more of an issue for data analysis than for design principles in cell biology.
A third issue seems to have been resolved recently. In [Reshef et al., 2013] it is proposed that mutual information does not satisfy the concept of equitability and an alternative, equitable, dependence measure related to mutual information was suggested, the maximal information coefficient (MIC). This equitability concept can intuitively be taken to mean that a dependence

measure classifies relationships with equal noise levels, equally. More recently though, in [Kinney and Atwal, 2014] it was shown that their definition of equitability was not correct mathematically. An alternative formalization of the concept was introduced which mutual information does happen to satisfy. We will now briefly review their results.

2.5.1 Self-equitability

Definition 2.5.1. A dependence measure $D[X;Y]$ is self-equitable if and only if it is symmetric and satisfies

$$D[X;Y] = D[f(X);Y] \tag{2.30}$$

whenever f is a deterministic function, X and Y are variables of any type, and $X \leftrightarrow f(X) \leftrightarrow Y$ forms a Markov chain.

This notion of self-equitability is more general (weaker) than the DPI. In the Appendix Theorem 3 Kinney and Atwal prove that DPI-satisfying measures are self-equitable.

Theorem 2.5.1. Every DPI-satisfying dependence measure $D[X;Y]$ is self-equitable.

Proof. If $X \leftrightarrow f(X) \leftrightarrow Y$ is a Markov chain, so is $f(X) \leftrightarrow X \leftrightarrow f(X) \leftrightarrow Y$. Extracting subchains, we see that for any DPI-satisfying measure $D[X;Y]$,

$$X \leftrightarrow f(X) \leftrightarrow Y \Rightarrow D[X;Y] \leq D[f(X);Y]$$
$$f(X) \leftrightarrow X \leftrightarrow Y \Rightarrow D[f(X);Y] \leq D[X;Y].$$

Therefore $D[X;Y] = D[f(X);Y]$. Since mutual information satisfies DPI, mutual information is self-equitable. □

Mutual information is not the only self-equitable information measure: in Theorem 4 Kinney and Atwal go on to prove that all so-called F-information measures (which includes mutual information) are self-equitable.

2.6 Overview

Now that we have seen the various properties of mutual information, a clear picture has emerged that mutual information is all a mathematician could want to judge the quality of communication processes. All of its properties: non-negativity, symmetry, self-equitability, relative simplicity, scope (discrete and continuous), interpretation in terms of uncertainty reduction and the number of distinguishable signal states and its relationship to other concepts such as KL-divergence set it apart from all other contenders. However, we must keep in mind that mutual information was originally developed with engineering in mind and not biology. As such, there are various questions mutual information is not equipped to answer that relate to the quality of signal transduction and similar processes. We discuss some views on this in Chapter 5.
For any results that we did not include here, the reader is referred to [Cover and Thomas, 1991], the standard in the field of information theory.

3 | Stochastic kinetics

This chapter serves as background material for the next chapter, where we consider the linear noise approximation in relation to two-component systems. In this chapter we introduce the necessary material on stochastic kinetics and the linear noise approximation in particular.

The deterministic description of kinetics makes use of mass action kinetics (introduced in the next section) whereas the stochastic approach often makes use of the chemical master equation which describes the evolution over time of the probability of having certain copy numbers of chemical species in the system. Chemical reactions that involve only a relatively small number of particles of a certain chemical species, as is often the case in signal transduction and gene expression, may be sensitive to noise; by noise we mean inherent fluctuations in the processes underlying the reactions, such as diffusion. When copy numbers are low, reactions occur so infrequently that fluctuations arise spontaneously, and these can be quite substantial. If these fluctuations are propagated through a signaling network, it may also lead to relatively high fluctuations in the concentration of the output molecule such as a transcription factor.

3.1 Deterministic kinetics

Cellular events happen through collisions of molecules due to diffusion, which is a random process. This random behaviour can behave quite like a deterministic process when the copy numbers of the molecules involved are large. So, deterministic kinetics arises as a limit for copy numbers of the stochastic kinetic description.

3.1.1 Mass action kinetics

The simplest chemical reactions have no reaction intermediates and are called elementary reactions. As an example, consider a molecule A transforming into a product B in one step,

$$A \to B. \tag{3.1}$$

Empirical studies have shown that the rates at which such reactions occur are roughly proportional to the product of the concentrations of the reactants involved. Thus we could approximate the example reaction's rate by saying $f = k \cdot A$. Here v is the reaction rate, k is the rate constant, and A is the concentration of the reactant.

A note on notation: in many texts, the concentration of a metabolite, X, is denoted using square brackets $[X]$ to distinguish it from X itself. To avoid unnecessary complexity in later equations, we will not make this distinction as it should be clear to the reader which is meant from the context.

Reaction (3.1) depicts that one molecule of A is transformed into one molecule of B; we say that the *stoichiometric coefficient* of A is -1 (1 molecule is lost), and for B it is $+1$ (one molecule is created). The rate of change of A is the reaction rate f times the stoichiometry coefficient, i.e. $\dot{A} = -f = -kA$.
More generally, for reactions with multiple reactants and products:

$$n_1 A_1 + n_2 A_2 + \ldots \to m_1 B_1 + m_2 B_2 \ldots$$

the reaction rate can be approximated by

$$f = k \prod_i A_i^{n_i}. \tag{3.2}$$

This rate law is referred to as the *law of mass-action* or *mass-action kinetics* (MAK).

3.1.2 ODE system for reaction networks

In (3.2) there are several stoichiometric coefficients. When considering a reaction network with multiple of such reactions we summarize the coefficients in a stoichiometry matrix $\boldsymbol{\mathcal{N}}$, where the (i, j) entry is the stoichiometric coefficient of reactant i in the reaction j. This matrix is of size $N \times R$ where N is the number of molecular species and R the number of reactions. In such a system, each species can partake in multiple reactions so that if we use $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_N)^T$ to denote the $M \times 1$ vector of macroscopic concentrations which are time-dependent, we have

$$\frac{\partial \phi_i(t)}{\partial t} = \sum_{j=1}^{R} \boldsymbol{\mathcal{N}}_{ij} f_j(\boldsymbol{\phi}).$$

Now letting the vector $\boldsymbol{f}(\boldsymbol{\phi})$ denote the vector-valued transition rate function of size $R \times 1$, the evolution of the concentrations follows the ODE system

$$\frac{\partial \boldsymbol{\phi}}{\partial t} = \boldsymbol{\mathcal{N}} \mathbf{f}(\boldsymbol{\phi}). \tag{3.3}$$

The macroscopic description is inappropriate for many systems of interest, particularly when the molecular copy numbers are low. In such cases we require a stochastic or mesoscopic description, which we will focus on in the next sections. Much more can be said on the subject of deterministic molecular kinetics, but this is enough for the coming material in this thesis. Heinrich's text is a good introduction for further details [Heinrich and Schuster, 1996].

3.2 Stochastic kinetics: master equations

Master equations (or ME for short) are used to describe the stochastic evolution over time of a system that can be in one of a countable number of states at any given time. The master equations are then a set of differential equations of the probability that the system occupies each different state. In the next sections, we will provide a straightforward derivation of the forward master equation for birth-death processes and its continuous state-space equivalent. For alternative derivations, we refer to Chapters 2 and 3 of [Goel and Richter-Dyn, 1974] & [Van Kampen, 2007]. We will also discuss the steady state distribution for the simplest process.

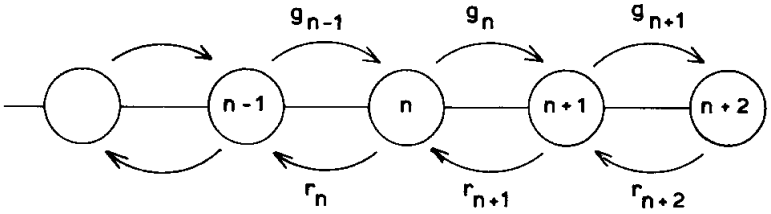


Figure 3.1: **Simple birth-death process.** A birth and death process jumps between neigboring states with transition probabilities g_n (jump to the right) and r_n (jump to the left).

3.2.1 Master equations for birth-death processes

A simple birth-death process

Assume that the stochastic process under consideration can only be in discrete states $\{\ldots, n-1, n, n+1, \ldots\}$ and assume that in each time interval of length Δt, the system moves from state n to state $n+1$ with probability g_n or moves to state $n-1$ with probability r_n; see Figure 3.1. These parameters may be constants or functions of n. Thus the system stays in state n with probability $1 - g_n - r_n$. In addition, assume for now that the system can only move over to neighboring states in this short interval. Then, writing $p_n(t)$ for the probability of being in state n at time t, we have that

$$p_n(t + \Delta t) = g_{n-1}p_{n-1}(t)\Delta t + (1 - g_n - r_n)p_n(t)\Delta t + r_{n+1}p_{n+1}(t)\Delta t. \tag{3.4}$$

Taylor expanding the left-hand side up to $O\left((\Delta t)^2\right)$, we find

$$\frac{\partial}{\partial t}p_n(t)\Delta t = p_n(t + \Delta t) - p_n(t) + O\left((\Delta t)^2\right),$$

so that up to second order

$$\frac{\partial}{\partial t}p_n(t) = g_{n-1}p_{n-1}(t) - (g_n + r_n)p_n(t) + r_{n+1}p_{n+1}(t). \tag{3.5}$$

Equation (3.5) is the forward master equation. Note that it is an ordinary differential equation in the time variable t, yet discrete in the state variable n.

General birth-death processes

In the birth and death process above, the system only jumps between neighboring states. However, there are systems that require jumps from any state to any other state. In such cases, the two parameters g and r have to be replaced by a matrix W of transition probabilities. Here element (i, j) would contain the transition probability for moving from state j to state i. Rewritten, the master equation becomes

$$\frac{\partial}{\partial t}p_n(t) = \sum_m \left[W_{m\to n}p_{n'}(t) - W_{n\to m}p_n(t)\right]. \tag{3.6}$$

This model may sometimes be referred to as a gain-loss equation.

Birth-death process with equal rates

On the other hand, we can also simplify (3.5) by setting $g_n = g$ and $r_n = rn$ for all n:

$$\frac{\partial}{\partial t}p_n(t) = gp_{n-1}(t) - (g + rn)p_n(t) + r(n+1)p_{n+1}(t). \tag{3.7}$$

Note the multiplication of r with n and $n+1$, stemming from the fact that when there are n molecules, they could all possibly be the next one to be degraded. We can simplify this further by re-scaling time $t = \tau/r$ so that in the new time variable

$$\begin{aligned}\frac{\partial}{\partial \tau}p_n(\tau) &= \frac{\partial}{\partial t}p_n(t)\frac{dt}{d\tau} \\ &= \frac{g}{r}p_{n-1}(\tau) - \frac{g}{r}p_n(\tau) - np_n(\tau) + (n+1)p_{n+1}(\tau).\end{aligned} \tag{3.8}$$

Finally, introducing the scaled parameter $\lambda = g/r$ we get

$$\frac{\partial}{\partial \tau}p_n(\tau) = \lambda\left(p_{n-1}(\tau) - p_n(\tau)\right) - np_n(\tau) + (n+1)p_{n+1}(\tau). \tag{3.9}$$

3.2.2 Steady-state distribution for a simple birth-death process

In addition to the various forms of generality in which we can mould the master equation, it is common in the literature to rewrite master equations (especially for more difficult systems) in terms of a step-operator $\mathcal{S}$ to keep formulas short and concise. The step-operator $\mathcal{S}$ satisfies the following property for functions f with discrete arguments n:

$$\mathcal{S}^k f(n) = f(n+k). \tag{3.10}$$

With this new operator, (3.5) can be rewritten as

$$\frac{\partial}{\partial t}p_n(t) = (\mathcal{S} - 1)r_n p_n(t) + (\mathcal{S}^{-1} - 1)g_n p_n(t). \tag{3.11}$$

And similarly, rewriting (3.9):

$$\frac{\partial}{\partial \tau}p_n(\tau) = (\mathcal{S} - 1)\left(np_n(\tau) - \lambda p_{n-1}(\tau)\right). \tag{3.12}$$

Using this step-operator notation, we can easily find the steady-state distribution of the master equation. Considering the equation above at steady-state, we see that the second term has to be zero. At steady state we thus find a recursive relationship for p_n

$$p_n = \frac{\lambda}{n}p_{n-1} \Rightarrow p_n = \frac{\lambda^n}{n!}p_0.$$

To find p_0 we normalize the distribution:

$$\begin{aligned}1 &= p_0\sum_{n=0}^{\infty}\frac{\lambda^n}{n!} \\ p_0 &= e^{-\lambda}.\end{aligned} \tag{3.13}$$

Therefore we have found that $p_n = \frac{\lambda^n}{n!}e^{-\lambda}$, i.e. p_n has a Poisson distribution at steady-state. How are we to interpret this result? The Poisson distribution is a limiting form of the binomial distribution for $n \to \infty$ and keeping np constant, meaning that it resembles a large number of Bernoulli trials, each with small probability of success for each trial. Also we know that the mean and variance of the Poisson distribution are equal to its parameter $\lambda = g/r$.

3.2.3 The master equation for reaction networks

Above we introduced reaction networks for which we can write down the deterministic description (3.3) which can be solved by linearization around the stationary solution with standard methods. Here we will consider the stochastic description of such a network which we will use for the linear noise approximation.
Consider a system of volume (or size) Ω and N chemical species related through R elementary reactions with transition rate $f_j(\mathbf{x})$. The concentrations of the species are summarized in the vector $\mathbf{x} = (x_1, \ldots, x_N)^T$. We write $\mathbf{X} = \Omega\mathbf{x}$ for the vector of copy numbers. Associated with the reaction network is the $N \times R$ stoichiometry matrix $\boldsymbol{\mathcal{N}}$ (similar to the macroscopic case we saw before), with entries $\boldsymbol{\mathcal{N}}_{ij}$ signifying the change in copy number for species i through reaction j.
In the limit of the system volume going to infinity at constant concentrations, the stochastic fluctuations become insignificant to the copy numbers of the species, making $\mathbf{x}$ and the transition rates $f_j(x, \Omega)$ deterministic. Following [Elf and Ehrenberg, 2003] we introduce the notation

$$\boldsymbol{\phi} = \lim_{\Omega\to\infty} \mathbf{x} \tag{3.14}$$

for the macroscopic concentration $\boldsymbol{\phi}$ and $\bar{\boldsymbol{\phi}}$ for the steady-state concentration. For a reaction network of the form we describe here, the master equation takes the complicated form [Van Kampen, 2007]

$$\frac{dP(\mathbf{X}, t)}{dt} = \Omega \sum_{j=1}^{R} \left(\prod_{i=1}^{N} \mathcal{S}^{-\boldsymbol{\mathcal{N}}_{ij}} - 1 \right) f_j(\mathbf{x}) P(\mathbf{X}, t). \tag{3.15}$$

The sum on the right-hand side is over all reactions j. Each term in this sum consists of two separate terms due to the $\prod_{i=1}^{N} \mathcal{S}^{-\boldsymbol{\mathcal{N}}_{ij}}$ and the -1. Terms due to -1 give the probability of moving away from the current state $\mathbf{X}$ due to reaction j. The terms with $\prod_{i=1}^{N} \mathcal{S}^{-\boldsymbol{\mathcal{N}}_{ij}}$ give the probability of moving to state $\mathbf{X}$ from a different state through reaction j. This can be seen by observing that the negative exponent removes from the target state precisely those molecules that would be created/consumed through reaction j.

3.2.4 On simulations

For more complex systems than the birth-death process we discussed above, a full solution of the master equation is often not possible analytically. In some cases, progress can be made by deriving an approximate equation which is valid when protein numbers are large and or noise is small and we will focus on such methods in the next section. Often though, numerical simulations are performed to investigate system behavior. There are many user-friendly tools available to perform these stochastic simulations, notably: COPASI [Pahle et al., 2012] and StochPy [Maarleveld et al., 2013]. These tools implement (variants of) the famous Gillespie algorithm [Gillespie, 1977] for sampling trajectories from the master equation distribution. In this thesis we will not resort to these software packages as we are interested not just in the numeric results but in understanding and analytically calculating the linear noise approximation for a specific system. We therefore perform all our calculations in *Mathematica*, which can facilitate such symbolic computation.

3.3 The Fokker-Planck equation

Historically, the Fokker-Planck equation was first used by Fokker and Planck separately to describe Brownian motion, which was later expanded on by Einstein. In one variable x the general

Fokker-Planck equation is given by

$$\frac{\partial f(x,t)}{\partial t} = \left[-\frac{\partial}{\partial x}v(x) + \frac{1}{2}\frac{\partial^2}{\partial x^2}D(x)\right] f(x,t) \tag{3.16}$$

$$= -\frac{\partial J(x,t)}{\partial x}, \tag{3.17}$$

where $f(x,t)$ is the density function of the process under consideration that may evolve over time, $J(x,t) = \left[v(x) - \frac{1}{2}\frac{\partial}{\partial x}D(x)\right] f(x,t)$ is the probability flux and $v(x)$ and $D(x)$ are referred to as the drift and diffusion coefficient, respectively. Mathematically, equation (3.17) is a linear second-order partial differential equation of parabolic type. A parabolic PDE is a PDE of the form

$$\alpha u_{xx} + 2\beta u_{xy} + \gamma u_{yy} + \delta u_x + \epsilon u_y + \zeta = 0$$

such that $\beta^2 - \alpha\gamma = 0$, which is obviously true here since β and γ are zero. Remembering the diffusion equation $\frac{\partial u}{\partial t} = d\Delta u$ which we often come across in mathematical biology, the FP equation is a diffusion equation with an extra first-order derivative.
Note that a Fokker-Planck equation is per definition linear in $f(x,t)$. Consequently, when one comes across the word "linear" in relation to Fokker-Planck equations, what is meant is that $v(x)$ is linear in x and D is constant.
In N variables $\mathbf{x} = (x_1, \ldots, x_N)^T$, the Fokker-Planck equation is given by

$$\frac{\partial f(\mathbf{x},t)}{\partial t} = \left[-\sum_{i=1}^{N}\frac{\partial}{\partial x_i}v_i(\mathbf{x}) + \frac{1}{2}\sum_{i,j=1}^{N}\frac{\partial^2}{\partial x_i \partial x_j}D_{ij}(\mathbf{x})\right] f(\mathbf{x},t). \tag{3.18}$$

The collections of $v_i(\mathbf{x})$ and $D_{ij}(\mathbf{x})$ are now referred to as the drift vector and diffusion tensor respectively. The special case of the linear multivariate Fokker-Planck equation is

$$\frac{\partial f(\mathbf{x},t)}{\partial t} = \left[-\sum_{i,j=1}^{N}\mathbf{A}_{ij}\frac{\partial}{\partial x_i}x_j + \frac{1}{2}\sum_{i,j=1}^{N}\mathbf{B}_{ij}\frac{\partial^2}{\partial x_i \partial x_j}\right] f(\mathbf{x},t), \tag{3.19}$$

where $\mathbf{A}$ and $\mathbf{B}$ are now constant matrices and additionally $\mathbf{B}$ has to be symmetric.
The Fokker-Planck equation comes up in relation to approximations of the master equation.

3.3.1 Fokker-Planck approximation of the master equation

Under the assumption that average copy numbers are large, $n \gg 1$, n can be approximated as a continuous variable. Consistent with our earlier notation we will now include n within the parentheses $p(n,t)$, $r(n)$ and $g(n)$ when treating n as a continuous variable as opposed to a subscript p_n when treating n as discrete. For continuous n, we can approximate the master equation with a Fokker-Planck equation. Consider (3.5) where we now transition to the continuous notation and Taylor expand the $g(n-1)p(n-1)$ and $r(n+1)p(n+1)$ terms as

$$g(n-1)p(n-1) = g(n)p(n) - \frac{\partial}{\partial n}[g(n)p(n)] + \frac{1}{2}\frac{\partial^2}{\partial n^2}[g(n)p(n)]$$
$$r(n+1)p(n+1) = r(n)p(n) + \frac{\partial}{\partial n}[r(n)p(n)] + \frac{1}{2}\frac{\partial^2}{\partial n^2}[r(n)p(n)].$$

The parameters g and r appear within the derivatives since they could be functions of n. Substituting this in (3.5), the first terms cancel so that

$$\begin{aligned}\frac{\partial}{\partial t}p(n) &= \frac{\partial}{\partial n}[r(n)p(n)] - \frac{\partial}{\partial n}[g(n)p(n)] + \frac{1}{2}\left[\frac{\partial^2}{\partial n^2}[g(n)p(n)] + \frac{\partial^2}{\partial n^2}[r(n)p(n)]\right] \\ &= -\frac{\partial}{\partial n}\left[(g(n) - r(n))\,p(n)\right] + \frac{1}{2}\frac{\partial^2}{\partial n^2}\left[(r(n) + g(n))\,p(n)\right].\end{aligned}$$

Defining $v(n) \equiv g(n) - r(n)$ and $D(n) \equiv g(n) + r(n)$, we recover the Fokker-Planck equation

$$\frac{\partial}{\partial t}p(n) = -\frac{\partial}{\partial n}\left[v(n)p(n)\right] + \frac{1}{2}\frac{\partial^2}{\partial n^2}\left[D(n)p(n)\right], \tag{3.20}$$

with coefficients $v(n)$ and $D(n)$.

3.3.2 Kramers-Moyal expansion

A simple generalization of the Fokker-Planck equation is one that also contains higher derivatives of x. This is termed the Kramers-Moyal expansion

$$\frac{\partial f(x,t)}{\partial t} = \left[\sum_{n=1}^{\infty}\left(-\frac{\partial}{\partial x}\right)^n D^{(n)}(x)\right] f(x,t), \tag{3.21}$$

where $D^{(n)}$ are the Kramers-Moyal coefficients. Consider this expansion in relation to a stochastic process in which the variable x can only take on discrete values $x_n = kn$, where $n = 1, \ldots, N$ and in which only transitions to nearest neighbors occur.
Using the expansion an interesting observation can be made when considering the master equation in (3.5). First, consider that in general

$$\begin{aligned}f(x \pm k) &= f(x) \pm kf'(x) \pm \frac{1}{2}k^2 f''(x) + \ldots \\ &= \exp(\pm kd/dx)f(x) \\ &= \left[\sum_{n=0}^{\infty}\frac{(\pm kd/dx)^n}{n!}\right] f(x).\end{aligned} \tag{3.22}$$

Using this and switching the notation from n to x in the master equation, we see

$$\begin{aligned}\frac{\partial}{\partial t}p_x(t) &= g_{x-k}p_{x-k}(t) - (g_x + r_x)p_x(t) + r_{x+k}p_{x+k}(t) \\ &= \left[\exp(-kd/dx) - 1\right] g_x p_x(t) + \left[\exp(kd/dx) - 1\right] r_x p_x(t).\end{aligned}$$

As a side note, realize that in the master equation we take $k = 1$, but to go general we keep the k notation. We can mold the equation above in the form of the Kramers-Moyal expansion

$$\frac{\partial}{\partial t}p_x(t) = \sum_{n=1}^{\infty}(-d/dx)^n D^{(n)}(x)p_x(t), \tag{3.23}$$

where

$$D^{(n)} = \frac{k^n}{n!}\left[g_x(t) + (-1)^n r_x(t)\right]. \tag{3.24}$$

If the difference k between the discrete steps becomes smaller, the Kramers-Moyal coefficients $D^{(n)}$ also become smaller and we may approximate $p_x(t)$ by truncating the expansion at some

finite value of n. For an actual system we cannot change k though because it is determined by inherent physics as in the case of molecule numbers. What we might be able to do is increase the size of the system. If we increase the size of the system by a factor K, i.e. $n = 1, \ldots, NK$, extensive quantities (quantities that are proportional to the size of the system) will also increase by this factor, i.e., $x = nk = Kx^*$. We get

$$\frac{\partial}{\partial t} p_{x^*}(t) = \sum_{n=1}^{\infty} (-d/dx^*)^n D^{(n)}(x^*) p_{x^*}(t),$$
$$D^{(n)} = \frac{(k/K)^n}{n!} \left[g_{x^*}(t) + (-1)^n r_{x^*}(t) \right]$$

Thus, by increasing the size of the system the Kramers-Moyal coefficients also decrease more rapidly in n. This is related to the $1/\Omega$ expansion by van Kampen which we will discuss in the next section.

3.3.3 Stationary solutions for the Fokker-Planck equation

In the case of a linear drift vector and a constant diffusion tensor, the (linear) FP equation can be solved resulting in Gaussian distributions for both the stationary and in-stationary solutions. This result also holds for time-dependent matrices. In the next chapter, when considering the linear noise approximation, we will require the stationary solution to a linear Fokker-Planck equation at steady state. In particular, we will also be interested in the stationary covariance matrix. There are multiple ways to prove that the stationary solution of a Fokker-Planck equation is a Gaussian distribution but all are quite involved. Below we provide two that also find the stationary covariance matrix, one method uses a detailed balance approach [Gardiner, 1985], and the other [Van Kampen, 2007] uses a Gaussian ansatz and then calculates the first two moments.

Method 1: Gaussian ansatz

We will consider a derivation where we use the ansatz that the solution is a Gaussian. Gaussian distributions are fully determined by their first two moments so we calculate them first. For the first moment (the expectation), multiply (3.19) with x_k and integrate over x[1]:

$$\int \frac{\partial f(\mathbf{x},t)}{\partial t} x_k d\mathbf{x} = \int \left[-\sum_{i,j=1}^{N} \mathbf{A}_{ij} \frac{\partial}{\partial x_i} x_j + \frac{1}{2} \sum_{i,j=1}^{N} \mathbf{B}_{ij} \frac{\partial^2}{\partial x_i \partial x_j} \right] f(\mathbf{x},t) x_k d\mathbf{x},$$
$$\frac{\partial}{\partial t}\langle x_k \rangle = -\int x_k \sum_{i,j=1}^{N} \mathbf{A}_{ij} \frac{\partial}{\partial x_i}(x_j f(\mathbf{x},t)) d\mathbf{x} + \frac{1}{2} \int x_k \sum_{i,j=1}^{N} \mathbf{B}_{ij} \frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x},t) d\mathbf{x},$$
$$= \sum_j \mathbf{A}_{kj} \langle x_j \rangle. \tag{3.25}$$

For the second moment, multiply (3.19) by $x_k x_l$ and integrate over x:

$$\frac{\partial}{\partial t}\langle x_k x_l \rangle = \int \left[-\sum_{i,j=1}^{N} \mathbf{A}_{ij} \frac{\partial}{\partial x_i} x_j + \frac{1}{2} \sum_{i,j=1}^{N} \mathbf{B}_{ij} \frac{\partial^2}{\partial x_i \partial x_j} \right] f(\mathbf{x},t) x_k x_l d\mathbf{x},$$
$$= \sum_i \mathbf{A}_{ki} \langle x_i x_l \rangle + \sum_j \mathbf{A}_{lj} \langle x_k x_j \rangle + \mathbf{B}_{kl}. \tag{3.26}$$

[1] I still have to understand the details of this calculation

In particular, we want the covariance matrix for the Gaussian solution which is related to the second moment by $\mathbf{\Sigma}_{kl} = \langle x_k x_l \rangle - \langle x_k \rangle \langle x_l \rangle$. Using the previous results, we see

$$\begin{aligned}\frac{\partial}{\partial t}\mathbf{\Sigma}_{kl} &= \frac{\partial}{\partial t}\langle x_k x_l \rangle - \frac{\partial}{\partial t}\langle x_k \rangle \frac{\partial}{\partial t}\langle x_l \rangle \\ \frac{\partial}{\partial t}\mathbf{\Sigma}_{kl} &= \sum_i \mathbf{A}_{ki}\langle x_i x_l \rangle + \sum_j \mathbf{A}_{lj}\langle x_k x_j \rangle - \sum_j \mathbf{A}_{kj}\langle x_j \rangle \sum_j \mathbf{A}_{lj}\langle x_j \rangle + \mathbf{B}_{kl} \end{aligned} \tag{3.27}$$

Rewriting in matrix form we see that $\mathbf{\Sigma}$ satisfies a Lyapunov equation

$$\frac{\partial}{\partial_t}\mathbf{\Sigma} = \mathbf{A}\mathbf{\Sigma} + \mathbf{\Sigma}\mathbf{A}^T + \mathbf{B}. \tag{3.28}$$

We conclude that the solution to (3.19) is the following Gaussian:

$$f_X(x,t) = (2\pi)^{r/2}(\det \mathbf{\Sigma})^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x - \langle x \rangle)\mathbf{\Sigma}^{-1}(x - \langle x \rangle)\right] \tag{3.29}$$

By substituting this in the Fokker-Planck equation, one can check that this really is a solution.

Method 2: Detailed balance on Ornstein-Uhlenbeck processes

We now look at a derivation of the Lyapunov equation at steady state that uses a detailed balance approach.
From (3.17) we see that the stationary solution to the Fokker-Planck equation requires the probability flux to vanish (equal zero). The concept of detailed balance is a generalization of this fact, where at the stationary state each transition must be balanced by its reverse transition.
In general, detailed balance is concerned with a set of variables x_i, that are transformed under time reversal, to the reversed variables $\epsilon_i x_i$ where $\epsilon_i = \pm 1$. These ϵ_i take care of whether x_i is odd or even under this time reversal.
The mathematical condition for detailed balance can then be understood to be

$$p_S(\mathbf{x}, t+\tau; \mathbf{x}', t) = p_S(\boldsymbol{\epsilon}\mathbf{x}', t+\tau; \boldsymbol{\epsilon}\mathbf{x}, t) \tag{3.30}$$

where $\mathbf{x}$ and $\boldsymbol{\epsilon}$ are the vectors of x_i and ϵ_i respectively and p_S is the stationary distribution of the process under consideration.
For the Fokker-Planck equation, the conditions for detailed balance can be written as (see [Gardiner, 1985] for the rather lengthy and complicated details):

$$\epsilon_i \epsilon_j B_{ij}(\boldsymbol{\epsilon}\mathbf{x}) = B_{ij}(\mathbf{x}) \tag{3.31}$$

$$\frac{1}{2}\left[A_i(\mathbf{x}) + \epsilon_i A_i(\boldsymbol{\epsilon}\mathbf{x})\right] - \frac{1}{2}\sum_j \frac{\partial}{\partial x_j} B_{ij}(\mathbf{x}) = -\frac{1}{2}\sum_j B_{ij}(\mathbf{x})\frac{\partial \phi(\mathbf{x})}{\partial x_j}. \tag{3.32}$$

For the linear Fokker-Planck equation we assume that $A_i(\mathbf{x}) = \sum_j A_{ij}x_j$, $B_{ij}(\mathbf{x}) = B_{ij}$ and $\phi(\mathbf{x}) = -\ln p_s(\mathbf{x})$ where $p_s(\mathbf{x})$ is the stationary distribution. The conditions then become

$$\epsilon_i \epsilon_j B_{ij} = B_{ij} \tag{3.33}$$

and

$$\begin{aligned}\sum_j A_{ij}x_j + \sum_j \epsilon_i \epsilon_j A_{ij} x_j - \sum_j \frac{\partial}{\partial x_j} B_{ij} &= \sum_j B_{ij}\frac{\partial}{\partial x_j}\log p_s(\mathbf{x}) \\ \sum_j \left[\epsilon_i \epsilon_j A_{ij} + A_{ij}\right] x_j &= \sum_j B_{ij}\frac{\partial}{\partial x_j}\log p_s(\mathbf{x}). \end{aligned} \tag{3.34}$$

Now notice that this second condition has a linear left-hand side so that the derivative of the stationary distribution must be linear in $\mathbf{x}$. This stationary distribution must therefore be Gaussian and in fact a Gaussian with mean zero since there is no constant term on the left. Thus we have that

$$p_s(\mathbf{x}) = (2\pi)^{-N/2} \left(\det \boldsymbol{\Sigma}\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{x}\boldsymbol{\Sigma}^{-1}\mathbf{x}^T\right).$$

Plugging this into (3.19) we see

$$0 = -\sum_i A_{ii} - \frac{1}{2}\sum_{i,j} B_{ij}\Sigma^{-1}_{ij} + \sum_{k,j}\left(\sum_i \Sigma^{-1}_{ki}A_{ij} + \frac{1}{2}\sum_{i,l}\Sigma^{-1}_{ki}B_{il}\Sigma^{-1}_{lj}\right)x_k x_j$$

Both the constant terms and the quadratic terms in $\mathbf{x}$ vanish under the following matrix condition

$$\boldsymbol{\Sigma}^{-1}\mathbf{A} + \mathbf{A}^T\boldsymbol{\Sigma}^{-1} = -\boldsymbol{\Sigma}^{-1}\mathbf{B}\boldsymbol{\Sigma}^{-1}. \tag{3.35}$$

Which can be rewritten as $\mathbf{A}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\mathbf{A}^T = -\mathbf{B}$ which is the Lyapunov equation once again.

3.3.4 Lyapunov equations

Equation (3.28) is a continuous-time Lyapunov equation and a special case of the Sylvester equation, which is in turn a special case of the continuous time algebraic Riccati equation (CARE) when $\mathbf{B} = 0$:

(a) Lyapunov equation : $$\mathbf{AX} + \mathbf{XA^T} + \mathbf{Q} = 0, \tag{3.36}$$

(b) Sylvester equation : $$\mathbf{AX} + \mathbf{XB} + \mathbf{Q} = 0, \tag{3.37}$$

(c) Ricatti equation : $$\mathbf{A}^T\mathbf{X} + \mathbf{XA} - \mathbf{XBR^{-1}B^TX} + \mathbf{Q} = 0. \tag{3.38}$$

The Bartels-Stewart algorithm [Bartels and Stewart, 1972] can be used to efficiently solve Lyapunov equations numerically. This is done by transforming $\mathbf{A}$ and $\mathbf{B}$ into Schur form ($\mathbf{A} = \mathbf{QUQ}^{-1}$) with the help of a QR decomposition. This results in a triangular system which can be solved through back-substitution. In the next chapter we will perform analytical and numerical calculations on Lyapunov equations in *Mathematica*, which has a dedicated function *LyapunovSolve* for this purpose. For matrices of the size we will consider straightforward use of the Solve[...] function is approximately 4x slower for numeric matrices. It will turn out that analytically, LyapunovSolve and Solve are not of much use unless a smart trick is used.
There are some useful remarks to be made about the structure of the Lyapunov equation. First of all the left-hand side of $\mathbf{AX} + \mathbf{XA^T} = -\mathbf{Q}$ is linear in $\mathbf{X}$, which means that the entries of $\mathbf{X}$, when thought of as a vector, are linearly transformed to another vector. This in effect means that we can rewrite this equation with $\mathbf{X}$ and $\mathbf{Q}$ as (column-wise) vectors. Following notation used in [Horn, 1994] we associate with each mxn matrix $\mathbf{A}$ the column vector

$$\text{vec } \mathbf{A} = (a_{11}, \ldots, a_{m1}, a_{12}, \ldots, a_{m2}, \ldots, a_{1n}, \ldots, a_{mn})^T. \tag{3.39}$$

Using this and the Kronecker-delta product, we rewrite the Lyapunov equation as

$$\left[(\mathbf{I}\otimes\mathbf{A}) + (\mathbf{A}\otimes\mathbf{I})\right] \text{ vec } \mathbf{X} = \text{vec } \mathbf{Q}. \tag{3.40}$$

To see this, remember from linear algebra that if $\mathbf{A}$ is an mxn matrix and $\mathbf{B}$ is a pxq matrix, then the Kronecker product (or direct product) $\mathbf{A}\otimes\mathbf{B}$ is the mpxnq block matrix:

$$\mathbf{A}\otimes\mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix}. \tag{3.41}$$

This reduces the Lyapunov equation to a simple $\mathbf{Ax} = \mathbf{b}$ situation and might help in quickly solving it.
There are many more mathematically interesting properties of Lyapunov equations. However, we will not be using these in this text. We refer to [Horn, 1994] for further details.

3.4 The linear Noise approximation

Even though we might be able to write down the master equation it is not straightforward to solve it, either analytically or numerically. Analytical results do exist for one-step master equations as we saw, as do numerical algorithms. However, for large systems even these simulations become ever more costly and infeasible. Therefore methods have been developed for quick characterizations of processes, most notably for our purposes the linear noise approximation (LNA), also known as the fluctuation-dissipation theorem (FDT) or Van Kampen's system size expansion.
In essence, the linear noise approximation is a method for quickly estimating the (co)variances of molecular species in the system without having to simulate trajectories. The advantage of the LNA is that it is very applicable because it only uses information about the stoichiometry of the system and the macroscopic reaction rates to solve a matrix equation. It is therefore less computationally demanding than simulating using the Gillespie algorithm. Original work on the LNA was done by Van Kampen in [Van Kampen, 2007], and generalized in [Elf and Ehrenberg, 2003].

3.4.1 Derivation

This section follows the derivation in [Elf and Ehrenberg, 2003] which is a generalization, both in dimensional terms and in scope, of the results in [Van Kampen, 2007]. We attempt to include more mathematical detail than the derivation in [Elf and Ehrenberg, 2003].
The key idea is to expand the master equation in powers of a small parameter so as to have an objective measure for the size of the terms that will appear, allowing us to drop negligible terms. Looking at (3.15) we choose Ω, since it defines the volume of the system and therefore also sets the scale for the stochastic fluctuations. We perform the expansion knowing that for large Ω the fluctuations will be small, relatively, which is consistent with the convergence of the stochastic and deterministic descriptions. Thus, since we wish to expand in a small parameter we will expand in *negative* powers of Ω.
Step one in the derivation is a change of variables that is aptly chosen. The copy number vector $\mathbf{X}$ is to be decomposed in a deterministic part $\Omega\boldsymbol{\phi}$ and a stochastic part depending on a stochastic variable $\boldsymbol{\xi}$. The key assumption that defines the system size expansion, is that we assume that the fluctuations scale with the square root of the system size/volume. To that end define the new variable $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_n)$ using the relations

$$\mathbf{X} = \Omega\boldsymbol{\phi} + \sqrt{\Omega}\boldsymbol{\xi} \implies \boldsymbol{\xi} = \frac{\mathbf{X} - \Omega\boldsymbol{\phi}}{\sqrt{\Omega}} \tag{3.42}$$

implying

$$\mathbf{x} = \boldsymbol{\phi} + \Omega^{-1/2}\boldsymbol{\xi} \implies \boldsymbol{\xi} = (\mathbf{x} - \boldsymbol{\phi})\sqrt{\Omega}, \tag{3.43}$$

where ϕ is as in (3.14). These new variables $\boldsymbol{\xi}$ are estimates for the fluctuations in $\mathbf{X}$, around the macroscopic concentrations in terms of the system size.
The second step in the derivation can be summarized as simply carrying through this change of variables and expanding the step operator, rate equation vector and probability distribution in powers of Ω.

Just as we have the probability distribution $P(\mathbf{X},t)$ for $\mathbf{X}$ we can now write $\Pi(\boldsymbol{\xi},t)$ for the probability distribution of $\boldsymbol{\xi}=(\xi_1,\ldots,\xi_N)$. These distributions are trivially related through

$$P(\mathbf{X},t)=P(\Omega\boldsymbol{\phi}+\Omega^{1/2}\boldsymbol{\xi},t)=\Pi(\boldsymbol{\xi},t).$$

We now expand this distribution in terms of Ω. We assume a steady-state, so that $\frac{d}{dt}\mathbf{X}=\mathbf{0}$, i.e. molecule numbers are constant and consequently $\frac{d\xi_i}{dt}=-\Omega^{1/2}\frac{\partial\phi_i}{\partial t}$. Differentiating $\Pi(\boldsymbol{\xi},t)$ with respect to time and using the chain-rule, we observe that

$$\begin{aligned}\frac{\partial P(\mathbf{X},t)}{\partial t}=\frac{\partial\Pi(\boldsymbol{\xi},t)}{\partial t}&=\frac{\partial\Pi(\boldsymbol{\xi},t)}{\partial t}\frac{dt}{dt}+\sum_{i=1}^{N}\frac{\partial\Pi(\boldsymbol{\xi},t)}{\partial\xi_i}\frac{\partial\xi_i}{\partial t}\\&=\frac{\partial\Pi(\boldsymbol{\xi},t)}{\partial t}-\Omega^{1/2}\sum_{i=1}^{N}\frac{\partial\Pi(\boldsymbol{\xi},t)}{\partial\xi_i}\frac{\partial\phi_i}{\partial t},\end{aligned}\tag{3.44}$$

where $\frac{\partial\phi_i}{\partial t}$ satisfies the macroscopic dynamics (3.3). This takes care of the left-hand side of (3.15). Next, we Taylor expand the transition rate $f_j(\mathbf{x})$ around the macroscopic value $f_j(\boldsymbol{\phi})$ resulting in

$$\begin{aligned}f_j(\mathbf{x})&=f_j(\mathbf{X}/\Omega)\\&=f_j(\boldsymbol{\phi}+\Omega^{-1/2}\boldsymbol{\xi})\\&=f_j(\boldsymbol{\phi})+\Omega^{-1/2}\sum_{i=1}^{N}\frac{\partial f_j(\boldsymbol{\phi})}{\partial\phi_i}\xi_i+O(\Omega^{-1}).\end{aligned}\tag{3.45}$$

This last equality shows that $f_j(\mathbf{x})$ differs from $f_j(\boldsymbol{\phi})$ by a term of order Ω^{-1}, which is of the order of single molecules [2]. We deduce that in the limit for large Ω there is no difference between the stochastic and deterministic description because then $f_j(\mathbf{x})=f_j(\boldsymbol{\phi})$.
The only term in the master equation left to estimate is the step-operator term $\prod_{i=1}^{N}\mathcal{S}^{-\boldsymbol{\mathcal{N}}_{ij}}$, this is done via a differential operator. First, recall that $\mathcal{S}^k f(X)=f(X+k)$ so that

$$\begin{aligned}\mathcal{S}^k f(X)&=f(X+k)\\&=\exp\left(k\partial/\partial X\right)f(X)\\&=\left[1+k\frac{\partial}{\partial X}+\frac{\partial^2}{\partial X^2}+\ldots\right]f(X).\end{aligned}\tag{3.46}$$

We have changed variables to $\boldsymbol{\xi}$ so that also $\mathcal{S}^k f(\boldsymbol{\xi})=f\left(\frac{(X+k)-\Omega\boldsymbol{\phi}}{\sqrt{\Omega}}\right)=f\left(\boldsymbol{\xi}+k\Omega^{-1/2}\right)$. Combining the two results, we get

$$\mathcal{S}^k=1+k\Omega^{-1/2}\frac{\partial}{\partial\boldsymbol{\xi}}+\frac{1}{2}\Omega^{-1}k^2\frac{\partial^2}{\partial\boldsymbol{\xi}^2}+O\left(\Omega^{-3/2}\right).\tag{3.47}$$

Taken in conjunction with the product in which $\mathcal{S}$ appears, we have

$$\begin{aligned}\prod_{i=1}^{N}\mathcal{S}^{-\boldsymbol{\mathcal{N}}_{ij}}&=1-\Omega^{-1/2}\sum_{i=1}^{N}\boldsymbol{\mathcal{N}}_{ij}\frac{\partial}{\partial\xi_i}+\Omega^{-1}\left[\sum_{\substack{i,k=1\\i\neq k}}^{N}\boldsymbol{\mathcal{N}}_{ij}\boldsymbol{\mathcal{N}}_{kj}\frac{\partial^2}{\partial\xi_i\xi_k}+\sum_{i=1}^{N}\frac{\boldsymbol{\mathcal{N}}_{ij}^2}{2}\frac{\partial^2}{\partial\xi_i^2}\right]+O\left(\Omega^{-3/2}\right)\\&\approx 1-\Omega^{-1/2}\sum_{i=1}^{N}\boldsymbol{\mathcal{N}}_{ij}\frac{\partial}{\partial\xi_i}+\frac{1}{2}\Omega^{-1}\sum_{i,k=1}^{N}\boldsymbol{\mathcal{N}}_{ij}\boldsymbol{\mathcal{N}}_{kj}\frac{\partial^2}{\partial\xi_i\xi_k}+O\left(\Omega^{-3/2}\right).\end{aligned}\tag{3.48}$$

[2] This is actually a rather subtle point.

To see this we have to do some bookkeeping on the multiplied terms. Note that if we multiply only the first term of each instance of $\mathcal{S}^{-\mathcal{N}_{ij}}$ as determined by (3.47) we get a zeroth order term, precisely 1. The only way of getting a $\Omega^{1/2}$ term is by multiplying one of the second terms with only ones. There will be N such terms, so that we get $-\Omega^{-1/2}\sum_i^N \boldsymbol{\mathcal{N}}_{ij}\frac{\partial}{\partial\xi_i}$. There are two ways of getting a term of order Ω^{-1}. Either all ones and two multiplications of the second term, or all ones and one of the third term. The two multiplications of the second terms cannot be for the same index in the sum. Note that the second way is the same as the first for $i = k$, if we disregard the $\frac{1}{2}$. In the second line we therefore combine these into one term and sum over all i and k. This is an approximation however, since we put the $\frac{1}{2}$ in front of the entire summation. Remembering the master-equation (3.15) and substituting the calculated expansions (3.44), (3.45) and (3.48), we have

$$\begin{aligned}
&\frac{\partial\Pi(\boldsymbol{\xi},t)}{\partial t} - \Omega^{1/2}\sum_{i=1}^{N}\frac{\partial\Pi(\boldsymbol{\xi},t)}{\partial\xi_i}\frac{\partial\phi_i}{\partial t} = \\
&\Omega\sum_{j=1}^{R}\left(-\Omega^{-1/2}\sum_{i=1}^{N}\boldsymbol{\mathcal{N}}_{ij}\frac{\partial}{\partial\xi_i} + \frac{1}{2}\Omega^{-1}\sum_{k=1}^{N}\sum_{l=1}^{N}\boldsymbol{\mathcal{N}}_{kj}\boldsymbol{\mathcal{N}}_{lj}\frac{\partial^2}{\partial\xi_k\xi_l} + O\left(\Omega^{-3/2}\right)\right) \\
&\times\left[f_j(\boldsymbol{\phi}) + \Omega^{-1/2}\sum_{m=1}^{N}\frac{\partial f_j(\boldsymbol{\phi})}{\partial\phi_m}\xi_m + O(\Omega^{-1})\right]\Pi(\boldsymbol{\xi},t).
\end{aligned} \tag{3.49}$$

Observe that we have renamed some indices to avoid confusion.
We wanted an expansion in negative powers of Ω but on both sides terms of order $\Omega^{1/2}$ remain:

$$-\Omega^{\frac{1}{2}}\sum_{i=1}^{N}\frac{\partial\Pi(\boldsymbol{\xi},t)}{\partial\xi_i}\frac{\partial\phi_i}{\partial t} = -\Omega^{1/2}\sum_{j=1}^{R}\sum_{i=1}^{N}\boldsymbol{\mathcal{N}}_{ij}f_j(\boldsymbol{\phi})\frac{\partial\Pi(\boldsymbol{\xi},t)}{\partial\xi_i}. \tag{3.50}$$

However, these terms are equal due to the macroscopic reaction $\frac{\partial\phi_i}{\partial t} = \sum_{j=1}^{R}\boldsymbol{\mathcal{N}}_{ij}f_j(\boldsymbol{\phi})$, which we saw in equation (3.3).
Collecting only the lowest order terms of Ω^0 we will find a particularly nice mathematical form. Identifying the right terms in (3.49) we see

$$\begin{aligned}
\frac{\partial\Pi(\boldsymbol{\xi},t)}{\partial t} &= \sum_{j=1}^{R}\left[-\sum_{i=1}^{N}\boldsymbol{\mathcal{N}}_{ij}\frac{\partial}{\partial\xi_i}\sum_{m=1}^{N}\frac{\partial f_j(\boldsymbol{\phi})}{\partial\phi_m}\xi_m + \frac{f_j(\boldsymbol{\phi})}{2}\sum_{k=1}^{N}\sum_{l=1}^{N}\boldsymbol{\mathcal{N}}_{kj}\boldsymbol{\mathcal{N}}_{ij}\frac{\partial^2}{\partial\xi_k\xi_l}\right]\Pi(\boldsymbol{\xi},t) \\
&= \sum_{j=1}^{R}\left[-\sum_{i=1}^{N}\boldsymbol{\mathcal{N}}_{ij}\sum_{m=1}^{N}\frac{\partial f_j(\boldsymbol{\phi})}{\partial\phi_m}\frac{\partial\left(\xi_m\Pi(\boldsymbol{\xi},t)\right)}{\partial\xi_i} + \frac{f_j(\boldsymbol{\phi})}{2}\sum_{k=1}^{N}\sum_{l=1}^{N}\boldsymbol{\mathcal{N}}_{kj}\boldsymbol{\mathcal{N}}_{lj}\frac{\partial^2\Pi(\boldsymbol{\xi},t)}{\partial\xi_k\xi_l}\right].
\end{aligned}$$

We can recover a linear Fokker-Planck equation with coefficient matrices $\mathbf{A}$ and $\mathbf{B}$ by rewriting the terms in front of the derivatives as

$$\mathbf{A}_{im} = \sum_{j=1}^{R}\boldsymbol{\mathcal{N}}_{ij}\frac{\partial f_j(\boldsymbol{\phi})}{\partial\phi_m} = \frac{\partial(\boldsymbol{\mathcal{N}}_{i\cdot}f(\boldsymbol{\phi}))}{\partial\phi_m} \tag{3.51}$$

$$\mathbf{B}_{kl} = \sum_{j=1}^{R}f_j(\boldsymbol{\phi})\boldsymbol{\mathcal{N}}_{kj}\boldsymbol{\mathcal{N}}_{lj} = \left[\boldsymbol{\mathcal{N}}\mathrm{diag}(f(\boldsymbol{\phi}))\boldsymbol{\mathcal{N}}^T\right]_{kl}. \tag{3.52}$$

Finally, we have

$$\frac{\partial \Pi(\boldsymbol{\xi}, t)}{\partial t} = -\sum_{i=1}^{N}\sum_{m=1}^{N} \mathbf{A}_{im} \frac{\partial (\xi_m \Pi(\boldsymbol{\xi}, t))}{\partial \xi_i} + \frac{1}{2}\sum_{k=1}^{N}\sum_{l=1}^{N} \mathbf{B}_{kl} \frac{\partial^2 \Pi(\boldsymbol{\xi}, t)}{\partial \xi_k \partial \xi_l}. \tag{3.53}$$

The stationary solution, ($\frac{\partial \Pi(\boldsymbol{\xi}, t)}{\partial t} = 0$) of (3.53) is a multivariate Gaussian distribution with the zero vector as mean (as we saw in an earlier section), i.e.

$$f(\boldsymbol{\xi}) = (2\pi)^{-N/2} (\det \boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\xi}^T \boldsymbol{\Sigma} \boldsymbol{\xi}\right). \tag{3.54}$$

where the covariance matrix $\boldsymbol{\Sigma}$ satisfies the Lyapunov equation (3.28)

$$\mathbf{A}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\mathbf{A}^T = -\mathbf{B}, \tag{3.55}$$

in which $\mathbf{A}$ and $\mathbf{B}$ are to be evaluated at the steady-state $\bar{\phi}$.

3.4.2 Discussion

The linear noise approximation thus states that, to lowest order, the stochastic fluctuations in a system like (3.3) are approximately normally distributed around the macroscopic mean with a certain covariance matrix $\boldsymbol{\Sigma}$. This approximation is valid in the large Ω limit, taken at constant concentrations. It follows that the linear noise approximation gives accurate results in the limit of large copy numbers and becomes less reliable for networks containing many species with low copy numbers of molecules. However, as noted before, gene regulation may involve small molecule numbers. Therefore, it is questionable how reliable the approximation is for these systems.
Recently, the Inverse Omega Square (IOS) method was implemented in a piece of open source software called iNA (intrinsic Noise Analyzer) [Thomas et al., 2013]. This method returns the variances and covariances of fluctuations about the means, calculated with the EMRE method, to an order accuracy higher than possible with the LNA. Therefore iNA provides a way to investigate the validity of the LNA for any biochemical network under study.
The fact that we take into account only zeroth order terms in powers of Ω and not the higher order terms in powers of $\Omega^{-\frac{1}{2}}$ and larger is why this approximation is frequently known as the *linear* noise approximation.
When applying the linear noise approximation, the trick is to find the covariance matrix from the $\mathbf{A}$ and $\mathbf{B}$ matrices which are easy to determine. As we noted in the previous chapter equation, (3.55) can be solved numerically and analytically and we will revisit both approaches when applying the linear noise approximation to a two-component system later on.

To summarize, three steps are essential to implementing the linear noise approximation to find the covariance matrix around a stationary steady state of a reaction network

1. Find the stoichiometry matrix $\boldsymbol{\mathcal{N}}$ and the stable equilibrium for the system,
2. Calculate $\mathbf{A}$ and $\mathbf{B}$ at steady state,
3. Finally, solve the Lyapunov equation for $\boldsymbol{\Sigma}$.

Steps 1 and 2 are usually quite easy. Even step 3 is easy, numerically, for relative large systems. However, analytically step 3 is generally hard or impossible. Only in two or three dimensional systems is it possible to make any reasonable analytic progress.
In the supplementary material of [Elf and Ehrenberg, 2003] some simple analytical examples are shown. For other examples, see [Hayot and Jayaprakash, 2004].

4 | Information processing in two-component systems

In this chapter we consider signal transduction through two-component systems in light of information theory (Chapter 2) and the mathematical machinery, especially the linear noise approximation, introduced in Chapter 3. First, we introduce the two-component signal transduction system. Then we will investigate some steady state properties followed by an application of the linear noise approximation. We conclude by comparing the results with recent literature [Maity et al., 2014].

4.1 Biology of two-component systems

> Two-component systems serve as a basic stimulus-response coupling mechanism to allow organisms to sense and respond to changes in many different environmental conditions. [Stock et al., 2000].

Prokaryotes use two-component signal transduction systems almost exclusively to sense and initiate responses to environmental conditions. Two-component systems rely on just two proteins: a sensor protein (S) and a response regulator protein (R), where the sensor is often membrane embedded and the response regulator is free in the cytosol and acts as a transcription factor. These systems are of importance in a large number of processes, such as chemotaxis and osmoregulation.
In detail the process can be described as follows: Binding of a ligand (signal), L, to the sensor S activates its autokinase activity. A kinase is a specific type of enzyme that can transfer phosphate groups from a molecule such as ATP, which gives up a phosphate group, to a substrate. This process is referred to as phosphorylation. In autokinase activity the kinase is itself the substrate. This process thus leads to a phosphorylated sensor (S_P). The phospate group can then be transferred to the response regulator R. This response regulator is often a transcription factor, which becomes activated by the phosphorylation step. The phosphorylated response regulator, R_P, can then go on to regulate transcription. The two-step process just described is found in almost all bacterial two-component systems. However, we will study a system with one more crucial property. Specifically we will assume that the phosphorylated response regulator can become dephosphorylated by the unphosphorylated sensor S. Thus the sensor S is a bifunctional enzyme. We call two-component systems with this property bifunctional systems, and monofunctional, when S does not have this property. In the latter case there is an independent

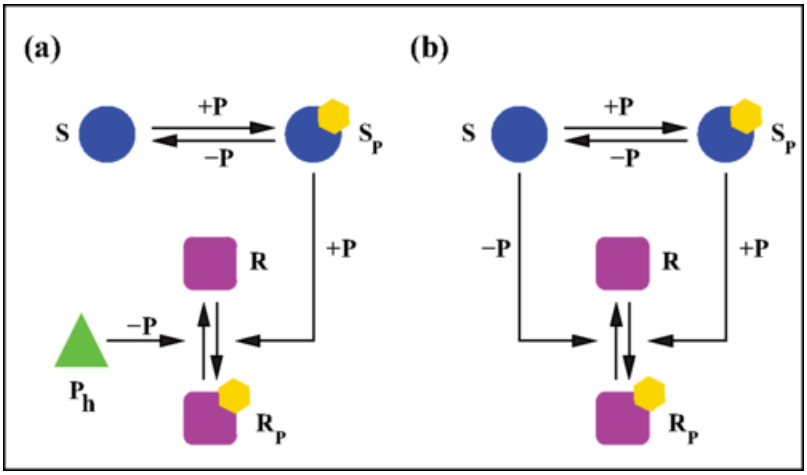


Figure 4.1: **Schematic of a two-component signal transduction system**. (a) a monofunctional TCS. (b) A bifunctional TCS. Ph stands for phosphatase. $\pm P$ stands for the addition or removal of the phosphate group (the orange hexagon). The crucial difference is that in a monofunctional TCS the sensor acts only as a source of the phosphate group, whereas, in a bifunctional TCS, it acts both as a source and sink for the phosphate group. Source: [Maity et al., 2014].

phosphotase in the system. Figure 4.1 displays a schematic of the process just described and indicates the difference between mono and bifunctional systems.
The only reaction types present in the general mechanism are protein complex formation (the joining of SP and R), phosphorylation (S to SP), dephosphorylation (RP back to R) and phosphotransfer (SP transferring a phosphate group to R to make RP). This is a surprisingly simplistic mechanism for such a widely used signaling system and raises the question how it functions so well. We will in particular be investigating how it functions in light of information transmission capabilities. First we will consider how to mathematically model the system and look at some intrinsic properties it displays.

4.2 Mathematical model of a two-component system

The schematic in Figure 4.1 can be modeled in various ways, depending on how much molecular detail we take into account. In full biological detail the bifunctional system looks like Figure 4.2. The system in Figure 4.2 has 13 reactions and 9 concentrations to keep track of and is thus not quite accommodating for analytic analysis. We would like to simplify this system by modeling it with less reactions and variables, and therefore less molecular detail, yet still capture the essentials so as to get useful results.
Note that all reactions contain an intermediate state in the full model by which we mean that if molecule A reacts with B to form C we say $A + B \to AB \to C$. For example, we have the sequence $S \to SL$, then $SL + P \to SLP$ and finally $SLP \to SP + L$, whereas we could simplify this part of the scheme by simply saying $S \to SP$ with L influencing the rate constant, as done in [Maity et al., 2014]. Also observe that all reactions in the complete model are reversible which is another possible avenue for simplification.
Let us examine this in light of recent literature. In [Maity et al., 2014] a TCS is considered without any complexes, but with full reversibility still in place. They consider only the molecular species:

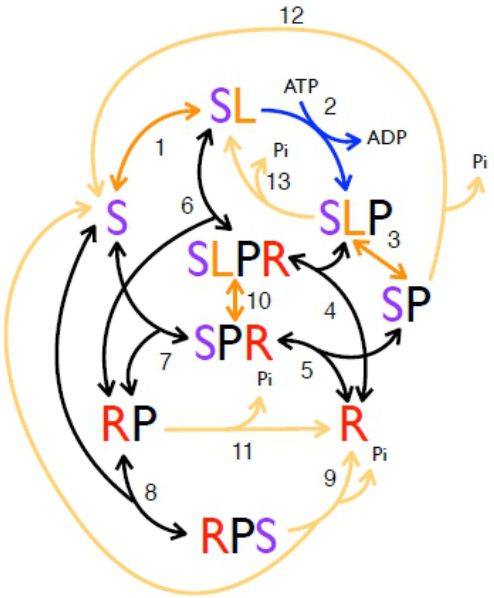


Figure 4.2: **Reactions in a full TCS model**. Here we depict a two-component system, now with emphasis on the reactions involved. There are 13 reactions in total. L is the signal molecule, S the sensor and R the response regulator. Source: F.J. Bruggeman.

R, R_P, S and S_P. A different TCS model is considered in [Shinar et al., 2007] where in addition, SL, R_PS and S_PR are considered. The latter is obviously more detailed than the former but still analytically tractable, as they show. We will consider the same species, however with a slight modification that results in a more structured stoichiometry matrix. More complex models have also been studied recently, for instance in unpublished work by Bruggeman en Maarleveld.

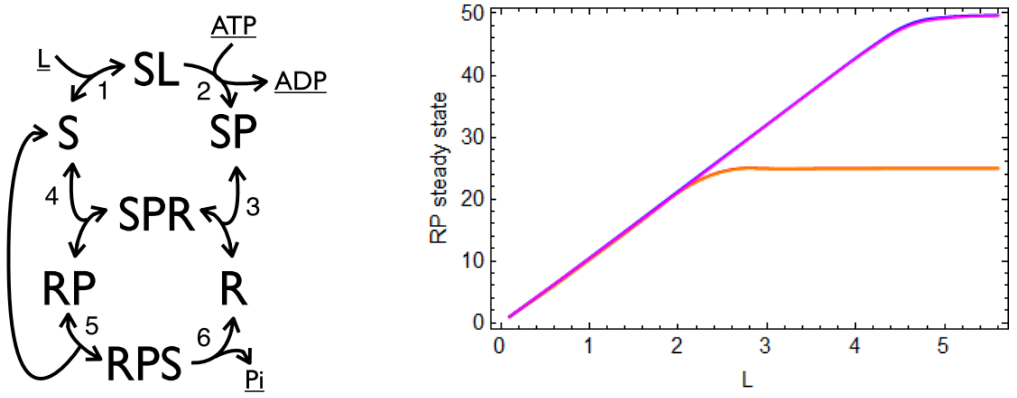
We consider the two-component system displayed in 4.3a, containing a subset of the reactions in 4.2:

$$S + L \underset{k_1^-}{\overset{k_1^+}{\rightleftharpoons}} SL + \text{ATP} \overset{k_2}{\rightarrow} S_P + L + \text{ADP} \tag{4.1}$$

$$S_P + R \underset{k_3^-}{\overset{k_3^+}{\rightleftharpoons}} S_PR \underset{k_4^-}{\overset{k_4^+}{\rightleftharpoons}} S + R_P \tag{4.2}$$

$$S + R_P \underset{k_5^-}{\overset{k_5^+}{\rightleftharpoons}} R_PS \overset{k_6}{\rightarrow} R + S + P_i. \tag{4.3}$$

Notice especially, that reaction 2 and reaction 6 are irreversible while the other reaction are reversible. We assume that the environmental signal L has a constant concentration, effectively making it a model parameter. Consequently, the model has 7 reactants and 6 reactions. P does not count as a variable in our system, we assume it to always be available or that its effect has been absorbed into a kinetic parameter. We impose mass-conservation of S and R: $S + SL + S_P + S_PR + R_PS = S_T$ and $R + R_P + R_PS + S_PR = R_T$. We can use these conservation relationships to exclude two species from our analysis which simplifies things somewhat.



(a) (b)

Figure 4.3: (a) **The two-component system we consider in this thesis**. This is a a reduced version of 4.2. Source: F.J. Bruggeman. (b) **Robustness**. The steady state concentration of R_P as a function of L as determined from numerical simulation. Two settings for R_T were used, 25 (orange) and 50 (magenta), nicely showing the bound by R_T as predicted by theory.

A particularly nice form of the stoichiometry matrix results from choosing to exclude S and R:

$$\mathbf{S} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}. \tag{4.4}$$

This banded, or bidiagonal, structure will carry through into the coefficient matrices of the Fokker-Planck equation $\mathbf{A}$ and $\mathbf{B}$ allowing us to use some linear algebra tricks later on.

Using mass-action kinetics (section 3.1) we turn the reactions into a system of equations (note that the derivatives are with respect to time but we have suppressed the time dependence in the right-hand side for readability)

$$\frac{\partial}{\partial t}SL = -k_1^- \cdot SL + k_1^+ \cdot L \cdot S - k_2 \cdot SL \tag{4.5}$$

$$\frac{\partial}{\partial t}S_P = k_2 \cdot SL + k_3^- \cdot S_P R - k_3^+ \cdot R \cdot S_P \tag{4.6}$$

$$\frac{\partial}{\partial t}S_P R = -k_3^- \cdot S_P R + k_3^+ \cdot R \cdot S_P + k_4^- \cdot R_P \cdot S - k_4^+ \cdot S_P R \tag{4.7}$$

$$\frac{\partial}{\partial t}R_P = -k_4^- \cdot R_P \cdot S + k_4^+ \cdot S_P R + k_5^- \cdot R_P S - k_5^+ \cdot R_P \cdot S \tag{4.8}$$

$$\frac{\partial}{\partial t}R_P S = -k_5^- \cdot R_P S + k_5^+ \cdot R_P \cdot S - k_6 \cdot R_P S. \tag{4.9}$$

Note that the only cross-product terms are $R \cdot S_P$ and $R_P \cdot S$ and that in fact, R_P and S_P do not arise outside of these two products at all.

4.3 Steady state analysis

A natural next step in analytically studying the two-component system is trying to find its analytic steady state. Having this steady state is also necessary to analytically calculate the linear noise approximation. It turns out that the system has two steady states one being a rather trivial steady state which exists for all parameter settings and a second steady state which requires there to be enough R to exist. First we consider a rather remarkable property that only bifunctional two-component systems display.

4.3.1 The Robustness property

Biological signaling systems can produce an output, such as in our case the concentration of a phosphorylated response regulator. The output concentration (R_P) as a function of the input concentration (L) is called the system's input-output relation. Because the system's component concentrations might vary over time and from cell to cell, one might expect that the input-output relation will also change. It is therefore interesting to consider in what way the input-output relationship is dependent on the concentrations of the system's components.
In [Shinar et al., 2007] it is shown that the two-component system we are investigating actually has the surprising property called robustness. With this we mean that the output of the system is not dependent on any components of the system. In fact it turns out that at steady state R_P is only linearly dependent on the signal L. In this section we show this result for the two-component system in Figure 4.3a following the derivation in [Shinar et al., 2007].
To start, consider Figure 4.3a and in particular note the way phosphate P goes in and out of the system. The influx of phosphate is via reaction 2 and the outflux is through reaction 6 giving rise to the mass balance

$$\frac{dp}{dt} = k_2 \mathrm{SL} - k_6 R_P S.$$

Now look at the mass balance for $R_P S$

$$\frac{d}{dt} R_P S = v_5 - v_6 = k_5^+ R_P \cdot S - k_5^- R_P S - k_6 R_P S.$$

Thus at steady state

$$(k_6 + k_5^-) R_P S = k_5^+ R_P \cdot S \tag{4.10}$$

$$R_P S = \frac{k_5^+}{k_6 + k_5^-} R_P \cdot S. \tag{4.11}$$

Substituting this steady state relationship for $R_P S$ into the mass balance for P at the steady state and solving for R_P, we see that

$$k_6 \frac{k_5^+}{k_6 + k_5^-} R_P \cdot S = k_2 \mathrm{SL}$$

$$R_P = \frac{k_2 (k_6 + k_5^-)}{k_6 k_5^+} \frac{\mathrm{SL}}{S}.$$

Now note that with the above result R_P is still dependent on S and thus not entirely robust. If we augment our argumentation above with the assumption that reaction 1 operates at thermodynamic equilibrium, i.e. we substitute $\mathrm{SL} = \frac{S \cdot L}{K_L}$ with $K_L = \frac{k_1^+}{k_1^- + k_2}$, then the S in the final result drops out and we have

$$R_P = \frac{k_2 (k_6 + k_5^-)}{k_6 k_5^+} \frac{\mathrm{L}}{K_L} \leq R_T. \tag{4.12}$$

This shows that the steady state concentration of the output system, R_P, is robust. In fact, it only depends in a linear fashion on the concentration of L and seven kinetic parameters. Since there is only a limited amount R_T of R available the steady state value of R_P will have to level off when it reaches R_T. We numerically simulated the model and plotted the steady state of R_P as a function of L in Figure 4.3b. As expected, it shows the linear input-output relationship, leveling of once R_P reaches the saturating level R_T.

4.3.2 Steady state and bifurcation behaviour

We now turn our attention to all steady-state concentrations in the system. The trivial steady steady state is just the zero vector which is an obvious solution to our system of equations. However, because we impose the constraints that there are positive amounts R_T and S_T, some concentrations must also be non-zero. Looking at the equations (4.9) we see that the only cross-product terms are $R \cdot S_P$ and $R_P \cdot S$ and that in fact, R_P and S_P do not arise on their own at all. Thus we see that S and R can go to zero while their phosphorylated counterparts do not and yet the system is still at steady state even if the complex concentrations are also zero. Using the conservation relationship we see that we must have $R_P = R_T$ and $S_P = S_T$.

The second steady state can be found by setting the left-hand side of (4.9) to zero, solving for all variables sequentially and carrying through. This is a nice calculation to do by hand and one will quickly find that the parameter combinations explode. The situation could perhaps be remedied somewhat by introducing smart parameter combinations as in [Shinar et al., 2007] however even they resort to expressing the steady state implicitly in terms of S and R. Remembering the conservation relationship we can express the steady state implicitly in terms of the R and S steady state values

$$\begin{pmatrix} SL^* \\ S_P^* \\ S_P R^* \\ R_P^* \\ R_P S^* \\ S^* \\ R^* \end{pmatrix} = \begin{pmatrix} 0 \\ S_T \\ 0 \\ R_T \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} SL^* \\ S_P^* \\ S_P R^* \\ R_P^* \\ R_P S^* \\ S^* \\ R^* \end{pmatrix} = \begin{pmatrix} \frac{k_1^+ \cdot L \cdot S^*}{k_1^- + k_2} \\ \frac{k_2}{k_6} \frac{k_3^- k_4^- k_5^- + k_4^+ k_5^+ k_6 + k3_b^- (k_4^- + k_5^+) k_6}{k_4^+ k_5^+ k_3^+ \cdot R} \cdot SL^* \\ \frac{k_2}{k_6} \frac{k_5^+ k_6 + k_4^- (k_5^- + k_6)}{k_4^+ k_5^+} \cdot SL^* \\ \frac{k_1^+}{k_1^- + k_2} \frac{k_2 (k_5^- + k_6)}{k_5^+ k_6} L \\ \frac{k_2}{k_6} \cdot SL^* \\ S_T - SL^* - S_P^* - R_P S^* - S_P R^* \\ R_T - R_P^* - R_P S^* - S_P R^* \end{pmatrix}. \quad (4.13)$$

The second steady-state contains the remarkable result we already found in the previous section, as of course it should. Note that aside from R_P being robust at steady-state the fact that its concentration at steady-state does not depend on anything but model parameters means that it there has to be enough R_T available for the system to even reach this state. Lastly, we remark that for $R_T < R_P^*$ the trivial fixed point is attracting and for $R_T > R_P^*$ the non-trivial fixed point is attracting while the trivial fixed point is repelling. Therefore, since the second steady state does not exist for all parameter values we have discovered that as we increase the amount of R_T the system undergoes a transcritical bifurcation at the steady state value of R_P.

Scaling the equations will most likely reveal more interesting details about this system of equations. However, since the main goal in this text is application of the system size expansion it has not been included here since, as shown below, the analytical treatment will not be very useful anyway, most likely even with scaled equations.

4.4 Model I: constant L concentration

We will now focus on applying the linear noise approximation to the system introduced above with the concentration of L as a model parameter. Writing down the stoichiometry matrix $\mathbf{S}$ and $\mathrm{diag}(f(\varphi))$ is straightforward and from them we trivially form $\mathbf{B}$ (see the section on LNA for the formulas):

$$\mathbf{B} = \begin{pmatrix} v_1 + v_2 & -v_2 & 0 & 0 & 0 \\ -v_2 & v_2 + v_3 & -v_3 & 0 & 0 \\ 0 & -v_3 & v_3 + v_4 & -v_4 & 0 \\ 0 & 0 & -v_4 & v_4 + v_5 & -v_5 \\ 0 & 0 & 0 & -v_5 & v_5 + v_6 \end{pmatrix}. \tag{4.14}$$

Note that at steady-state all v_i are equal so that we can rewrite $\mathbf{B}$ as

$$\mathbf{B} = v_1 \begin{pmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{pmatrix}. \tag{4.15}$$

For $\mathbf{A}$ we have to work a little. by plugging in the reaction rates with mass-action kinetics and differentiating we have

$$\mathbf{A} = \begin{pmatrix} -(k_1^- + k_2) & 0 & 0 & 0 & 0 \\ k_2 & -k_3^+ \cdot R & k_3^- & 0 & 0 \\ 0 & k_3^+ \cdot R & -(k_3^- + k_4^+) & k_4^- \cdot S & 0 \\ 0 & 0 & k_4^+ & -(k_4^- \cdot S + k_5^+ \cdot S) & k_5^- \\ 0 & 0 & 0 & k_5^+ \cdot S & -(k_5^- + k_6) \end{pmatrix}.$$

Especially notice the tridiagonal form of $\mathbf{A}$. Changing the model to include different reactions or complexes changes $\mathbf{A}$ and also complicates the structure of $\mathbf{A}$. Therefore the model depicted in Figure 4.9 achieves an optimal balance between the possibility for analytic insight and biological reality.

4.4.1 Solving for the covariance matrix

The linear noise approximation now states that the covariance matrix satisfies $\mathbf{A\Sigma} + \mathbf{\Sigma A}^T = -\Omega\mathbf{B}$. Numerically it is quite straightforward to solve this for $\mathbf{\Sigma}$, but analytically it requires significant computational resources. We performed the calculations described below in *Mathematica*.

Numerically
Numerically performing the linear noise approximation on our two-component system is quite simple. We simply set up the reactions using mass-action kinetics, define the moiety for R and S and calculate the numerical steady-state, making sure that the parameters are such that we are in the non-trivial steady-state (enough R_T). We plug the steady-state and the parameters into the $\mathbf{A}$ and $\mathbf{B}$ matrices (making them numeric) and solve the Lyapunov equation for $\mathbf{\Sigma}$. Through a less trivial addition we can repeat this for changing parameters sets and investigate how entries in $\mathbf{\Sigma}$ change as a result; see the results in a later section.

Analytically

In investigating $\mathbf{\Sigma}$ analytically we have tried multiple approaches outlined below. Our main goal was to find the variance of R_P at steady state analytically in terms of the model parameters.
The most obvious route to take is to simply set up the model and the linear noise approximation in Mathematica and then use Solve[...] on the Lyapunov equation. This solution step however takes at least 8+ hours to converge to an answer. We need more sophisticated techniques to speed up calculation time.
A first line of investigation is to follow [Elf and Ehrenberg, 2003] and look at transformed variables (normal modes $\delta\tilde{X} = \mathbf{Q}^{-1}\delta X$ where $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$) where the Lyapunov equation changes and a nice expression can be found for $\mathbf{\Sigma}$ which utilizes the eigenvalues of $\mathbf{A}$. For a system larger than 3x3 this is problematic. From the fact that the diagonal entries are negative yet the columns of $\mathbf{A}$ either sum up to zero (columns 2-4) or sum up to a negative number (columns 1 and 5) we can deduce using the Gershgorin circle theorem (A.3.1) that $\mathbf{A}$ has non-positive eigenvalues. Beyond this we were unable to to say anything about the eigenvalues of $\mathbf{A}$ since solving the characteristic polynomial (which is of fifth degree) is not possible.
As a next approach, look at the structure of the Lyapunov equation and remember (3.40). In our case the coefficient matrix $[(\mathbf{I}\otimes\mathbf{A}) + (\mathbf{A}\otimes\mathbf{I})]$ is very sparse (coming from the sparseness of $\mathbf{A}$ and the identity matrix). This matrix is too big to be displayed here (25x25) but it has a banded tridiagonal structure. Running a solve routine such on this equation still takes several hours at least.
We found one way to reduce computation time and to get more insight: by performing an LU decomposition on the $\mathbf{A}$ matrix and then solve the Lyapunov equation for the covariance matrix $\mathbf{\Sigma}$ in two steps. For reasons we do not fully understand this forces a dramatic decrease in computation time.

LU decomposition of A

For tridiagonal matrices, like A, there exists a simple LU factorization algorithm related to Taylor's algorithm (see section A.4). Simply following this algorithm (which is doable by hand) results in

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{k_2}{-k_1^- - k_2} & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix} \tag{4.16}$$

$$\mathbf{U} = \begin{pmatrix} -k_1^- - k_2 & 0 & 0 & 0 & 0 \\ 0 & -k_3^+ \cdot R & k_3^- & 0 & 0 \\ 0 & 0 & -k_4^+ & k_4^- \cdot S & 0 \\ 0 & 0 & 0 & -k_5^+ \cdot S & k_5^- \\ 0 & 0 & 0 & 0 & -k_6 \end{pmatrix}. \tag{4.17}$$

The key step now is to rewrite the left-hand side of the Lyapunov equation as

$$\begin{aligned} \mathbf{A\Sigma} + \mathbf{\Sigma A}^T &= \mathbf{A\Sigma} + (\mathbf{A\Sigma})^T && \text{since } \mathbf{\Sigma} \text{ is symmetric} \\ &= \mathbf{LY} + (\mathbf{LY})^T && \text{call } \mathbf{U\Sigma} := \mathbf{Y}. \\ &= \mathbf{LY} + \mathbf{Y}^T\mathbf{L}^T. && \end{aligned} \tag{4.18}$$

This allows us to solve for $\mathbf{\Sigma}$ in two steps. In the first step we solve $\mathbf{LY} + \mathbf{Y}^T\mathbf{L}^T = -\mathbf{B}$ for $\mathbf{Y}$. In the second step we solve we solve $\mathbf{U\Sigma} = \mathbf{Y}$ for $\mathbf{\Sigma}$. These matrix equations are solved in a matter

of seconds in *Mathematica*. However, the resulting covariance matrix is very complex and takes almost 100 megabytes to save to disk in plain text. Of course we are mainly interested in the (4,4) entry but even this entry is 2,5 Mb large. Note that up to this point we have not even plugged in the analytic steady-state yet. This results in an even more gigantic result.
It is to be expected that there are many more hidden structures inside the Lyapunov equation that we could take advantage of. However, it is also to be expected that this will not result in more insight that we can get by numerical simulation. We conclude that our analytic results are not very insightful or useful. One could say this was to be expected for a model of this level of complexity but it provided an interesting learning experience nonetheless.

4.4.2 Parameter dependency of Var(Rp)

Even though analytic results prove difficult to come by and non-insightful we can consider some interesting questions by considering the analytic result numerically or through purely numerical calculations. For instance, what is the lowest amplitude of stochastic fluctuations in the output variable possible for given parameter ranges? Which reactions play major roles in controlling these fluctuations and which reactions don't? Motivated by these questions we now plot the variance of R_P as a function of the model parameters.
To investigate how the variance of R_p depends on the various model parameters we can simply set all parameters but the one under inspection to some numeric value in our analytic result (or use a purely numeric approach) and plot the remaining function in terms of that parameter. Note though that when that one parameter changes the steady-state values also change with it which has to accounted for. The resulting dependence plots we found are displayed in Figure 4.4 for two parameter sets. The first set of parameters was chosen rather arbitrarily, the second was inspired by kinetic parameter values reported in [Igoshin et al., 2008].
As can be seen from the Figure 4.4, the difference in results between the two parameter sets is large. In general notice that for all reversible reactions the corresponding forward and backward kinetic parameters influence the variance in opposing ways. Also, reaction 3 appears to have little influence of note on the variance of R_P. In addition k_4^+ and S_T also seem to have little effect. Perhaps most intriguing is the fact that the variance increases linearly with L. As the steady state level of R_P also increases linearly with L due to the robustness property, this could be validation for the seemingly general principle that fluctuations scale with mean protein levels [Bar-Even et al., 2006].

4.5 Model II: including dynamics of L

The model we have considered so far has assumed a constant level of L. Next, we consider an extension where we take into account dynamics of the L concentration. A simple way to make L dynamic is to assume a constant synthesis rate C and a linear degradation with rate D:

$$\dot{L} = C - D \cdot L. \tag{4.19}$$

This equation models the arrival and departure of L molecules to and from the receptor pocket. This model is crucially different from the constant L model though because the steady state of L is now coupled to the steady state of S. Consequently the steady state of R_P will also depend on S and through S on the rest of the concentrations. Therefore the robustness property has been lost. However, as shows below, the linear input-output relationship is maintained. For the new system we find the following steady-state which is structurally the same as before but now with

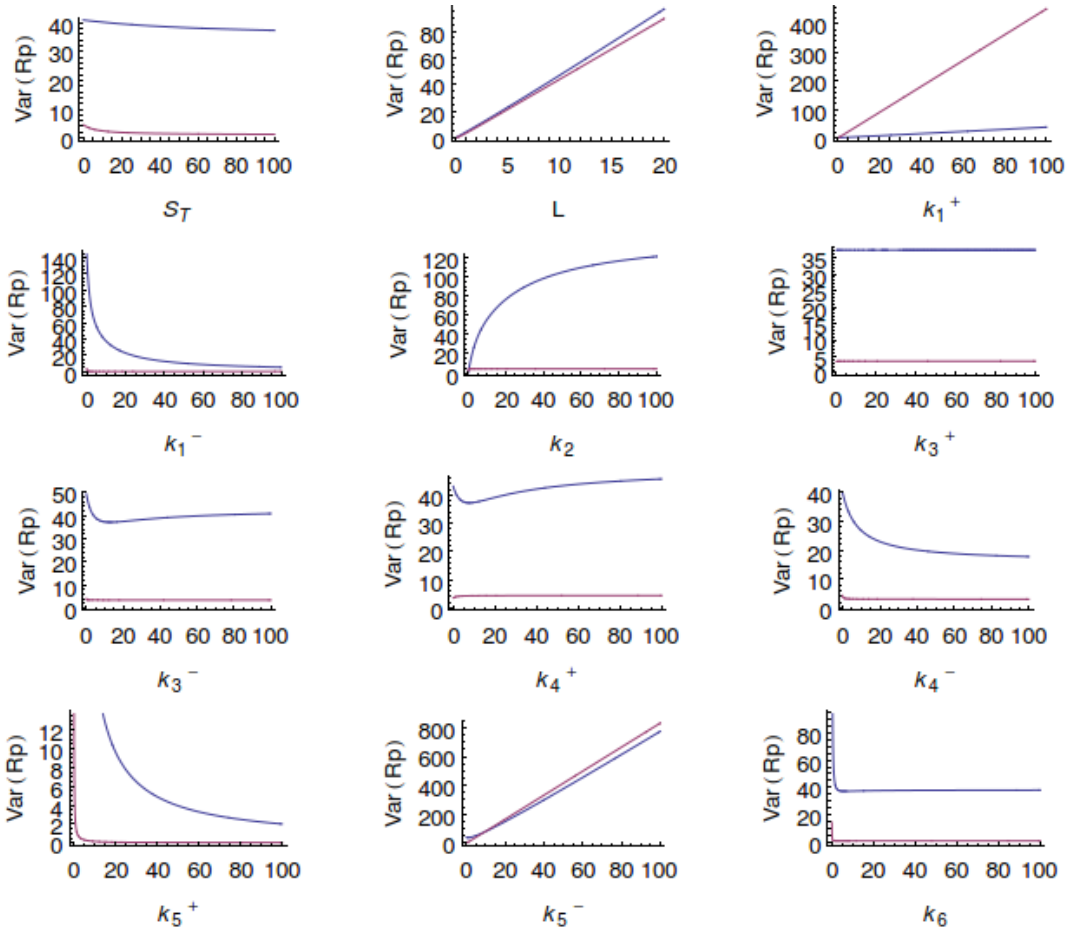


Figure 4.4: **Dependence of Var(Rp) on parameters.** Plots of the variance of R_p as analytically deduced using the linear noise approximation as a function of all model parameters. We take the analytical result for the Var(R_P) entry of the covariance matrix, plug in the analytical steady state and then plug in all but one parameter. We then plot the result in terms of the one remaining parameter. We did this for two parameter sets so as to show the influence of changing them. In blue: $R_T = 300, S_T = 100, L = 8, k_1^+ = 100, k_1^- = 10, k_2 = 5, k_3^+ = 1, k_3^- = 10, k_4^+ = 10, k_4^- = 1, k_5^+ = 5, k_5^- = 1, k_6 = 5$. In red: $R_T = 6, S_T = 3, L = 1, k_1^+ = 1, k_1^- = 0.01, k_2 = 0.1, k_3^+ = 1, k_3^- = 0.5, k_4^+ = 0.2, k_4^- = 0.5, k_5^+ = 0.5, k_5^- = 0.5, k_6 = 0.2$. This second set is an attempt at biologically realistic parameter values as inspired from [Igoshin et al., 2008].

an L steady state, again denoted implicitly in S and R,

$$\begin{pmatrix} L^* \\ SL^* \\ S_P^* \\ S_P R^* \\ R_P^* \\ R_P S^* \\ S^* \\ R^* \end{pmatrix} = \begin{pmatrix} \frac{C(k_1^- + k_2)}{D(k_1^- + k_2) + k_1^+ k_2 S^*} \\ \frac{k_1^+ \cdot L^* \cdot S}{k_1^- + k_2} \\ \frac{k_2}{k_6} \frac{k_3^- k_4^- k_5^- + k_4^+ k_5^+ k_6 + k3_b^- (k_4^- + k_5^+) k_6}{k_4^+ k_5^+ k_3^+ \cdot R} \cdot SL^* \\ \frac{k_2}{k_6} \frac{k_5^+ k_6 + k_4^- (k_5^- + k_6)}{k_4^+ k_5^+} \cdot SL^* \\ \frac{k_1^+}{k_1^- + k_2} \frac{k_2 (k_5^- + k_6)}{k_5^+ k_6} L^* \\ \frac{k_2}{k_6} \cdot SL^* \\ S_T - SL^* - S_P^* - R_P S^* - S_P R^* \\ R_T - R_P^* - R_P S^* - S_P R^* \end{pmatrix}. \tag{4.20}$$

Adding dynamics of L has only small consequences for the matrices in the linear noise approximation. The stoichiometry matrix $\mathbf{S}$ grows to be 6x7 but maintains the same structure. $\mathbf{A}$ and $\mathbf{B}$ also maintain the same structure. For example $\mathbf{A}$ becomes .

$$\mathbf{A} = \begin{pmatrix} -(D + k_1^+ \cdot S) & k_1^- & 0 & 0 & 0 & 0 \\ k_1^+ \cdot S & -(k_1^- + k_2) & 0 & 0 & 0 & 0 \\ 0 & k_2 & -k_3^+ \cdot R & k_3^- & 0 & 0 \\ 0 & 0 & k_3^+ \cdot R & -(k_3^- + k_4^+) & k_4^- \cdot S & 0 \\ 0 & 0 & 0 & k_4^+ & -(k_4^- \cdot S + k_5^+ \cdot S) & k_5^- \\ 0 & 0 & 0 & 0 & k_5^+ \cdot S & -(k_5^- + k_6) \end{pmatrix}.$$

Observe that $\mathbf{A}$ also maintains its tridiagonal structure. Our strategy for finding $\mathbf{\Sigma}$ analytically would thus remain the same. However, as it already produced huge results for the previous model one should not be hopeful of the outcome. In fact, it turns out that repetition of the LU-decomposition argument on this model has a vastly longer calculation time, so that we abandoned this approach and turned to numerical calculations instead.

Parameter dependence of Var(Rp)

In Figure 4.5 we plot the dependence of the variance of R_P on model parameters for the model including L dynamics. The chosen parameter values do not originate from any biologically inspired reasoning. The problem with this dynamic L model is that we do not yet understand what parameter sets bring the system to the non-trivial or the trivial steady state. This remains an avenue for further research. As such, the currently chosen values are simply chosen such that the non-trivial steady state is robust enough to allow the plots to be made.

Note that reaction 3 appears to have no influence of note on the variance of R_P. Second, for all reversible reactions the corresponding forward and backward kinetic parameters influence the variance in opposing ways. These results agree with the constant L model.

Again, the point of such plots is to be able to investigate which reactions contribute significantly or insignificantly to the variance in the output so that we can understand what drives and controls this variance.

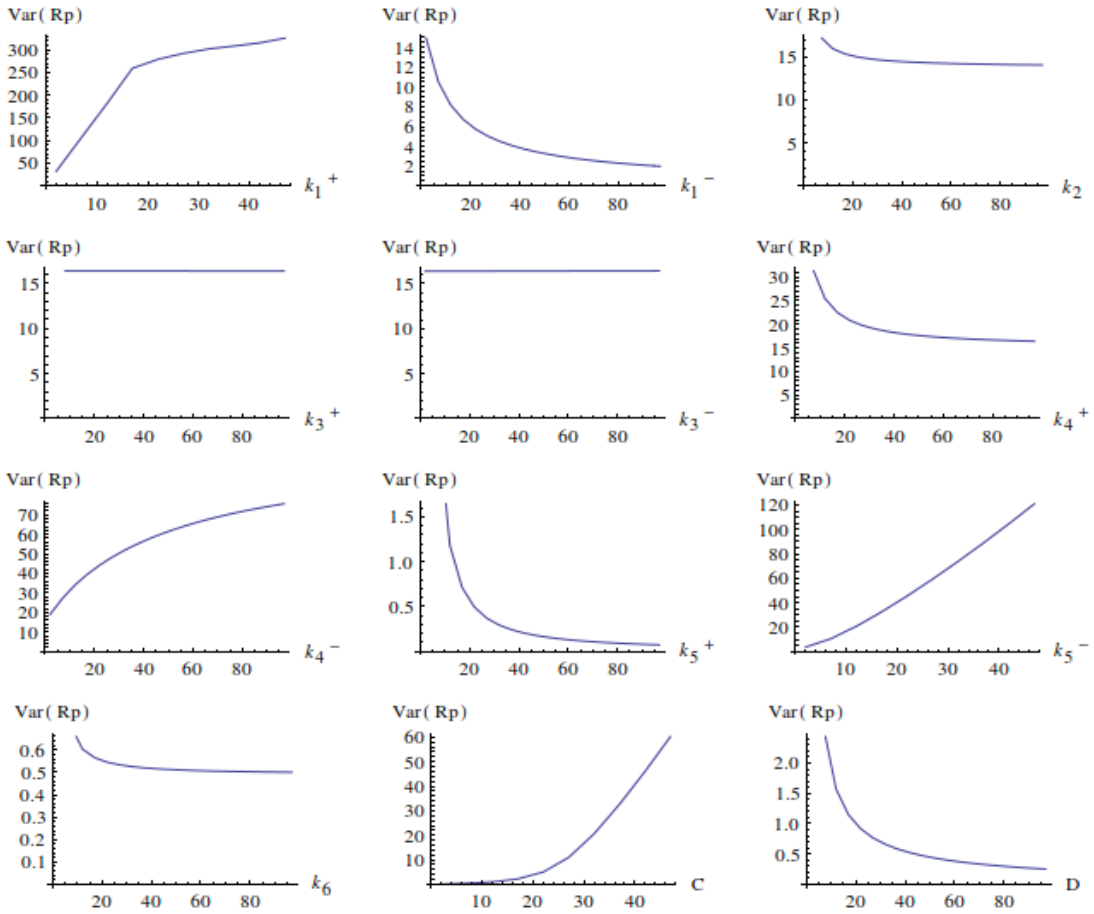


Figure 4.5: **Dependence of Var(Rp) on parameters.** Plots of the variance of R_p as numerically deduced using the linear noise approximation as a function of all model parameters. In each plot all variables except the one on the x-axis are kept constant. The fixed parameter values used in this plot are: $R_T = 400, S_T = 30, k_1^+ = 1, k_1^- = 1, k_2 = 10, k_3^+ = 10, k_3^- = 0.5, k_4^+ = 100, k_4^- = 0.5, k_5^+ = 2, k_5^- = 10, k_6 = 1, C = 30, D = 1$

4.6 Mutual information between L and Rp in a Gaussian channel

From the linear noise approximation we deduce that R_P has a distribution around the macroscopic steady state that is approximately Gaussian. The Lyapunov equation gives us the variance of this distribution and the covariance between L and R_P so we can now infer the mutual information between L and R_P. In the model including L dynamics the linear noise approximation states that L also has a Gaussian distribution around the steady state. Therefore we have a Gaussian channel for which equations (2.28), (A.3) hold.
It is problematic to actually use the $MI(L;R_P) = \frac{1}{2}\log_2(1+\text{SNR})$ result since it involves the gain parameter g which is not obvious here. It is much easier to use $MI(L;R_P) = -\frac{1}{2}\log_2(1-\rho^2)$ result since the covariance matrix gives us what we need to calculate ρ. Alternatively, we simply use the definition of mutual information in terms of integrals to do the calculation, taking what we need from the covariance matrix to estimate variances and plugging in the means that we plotted in the input-output relationship plots. Both methods return the same results, as they should. In the next section we show the actual results of this calculation along with the dependence of the steady state R_P level, the variance of R_P and the coefficient of variation on L at steady state.

4.7 Comparison with recent literature

A recently published paper ([Maity et al., 2014]) also investigated two-component signaling systems using analytic mathematical methods. Their model of the two-component system included no complex formation however, and it is therefore less detailed than the model we considered above. The assumption to neglect complex formation is a shaky one. It implicitly assumes that the concentrations of these complexes are so low that they are negligible which is not usually the case. Perhaps unsurprising, they made predictions about the mutual information and variance of R_P in terms of the steady state signal level that do not agree with our results.

4.7.1 Introduction to the Langevin noise approach

[Maity et al., 2014] use the Langevin noise approach instead of directly turning to Lyapunov equation as we did. In the Langevin approach one adds a stochastic Langevin noise term to each of the deterministic equations describing the system. This approach is mathematically equivalent to using the Fokker-Planck equation [Gardiner, 1985] but the Langevin approach is a little more concrete and therefore quite popular in the physical sciences. the This Langevin noise term $\xi_i(t)$ satisfies the following properties:

1. Its average vanishes: $\langle \xi_i(t) \rangle = 0$.
   This average is to interpreted as taken over an entire time series.

2. It has the following sharply peaked autocorrelation: $\langle \xi_i(t), \xi_i(t+\tau) \rangle = c \cdot \delta(\tau)$.
   Here δ is the Dirac delta function and c is a constant that has to be appropriately chosen so that it correctly reproduces the mean squared fluctuations at steady state.

Actually, the Langevin and Fokker-Planck equation approaches are only equivalent in case of the additional assumption that ξ has a Gaussian distribution with the above properties, again see [Gardiner, 1985] for the details.

Equations containing such Langevin terms are referred to as Langevin equations and are stochastic differential equations. To illustrate, in the two-component system the Langevin equation for L would be

$$\frac{dL}{dt} = C - D \cdot L + \xi_L.$$

After writing the system in terms of these Langevin equations, [Maity et al., 2014] go on to linearize the system and subsequently use Fourier transformations to obtain analytical expressions for the variance of R_P. The results they report can now be compared with our results.

4.7.2 Comparison of results

We imitated the 4 panel plot from their [Maity et al., 2014] for our model in Figure 4.6. We reproduced their results in Figure 4.7.

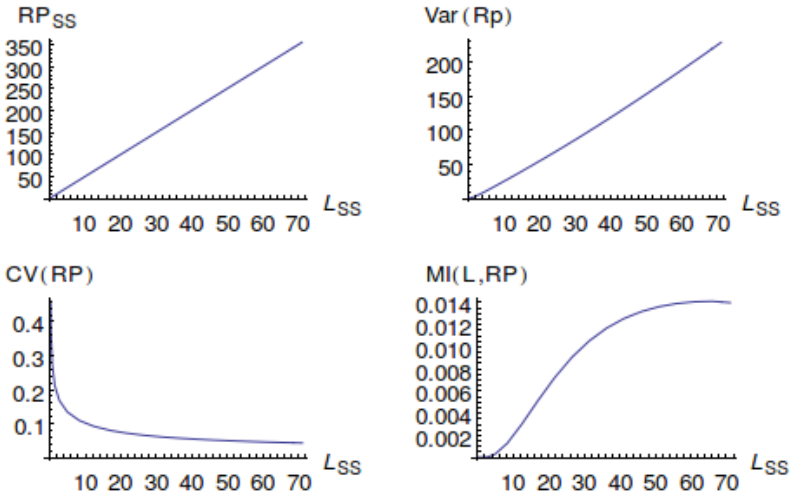


Figure 4.6: **Our results for the dynamic L model** (a) Steady state R_p level, (b) steady state variance $\sigma^2_{R_P}$, (c) steady state coefficient of variation $\sigma_{R_P}/\langle R_P \rangle$ and (d) steady state mutual information $I(I, R_p)$ as a function of mean extra-cellular signal level.

Consider Figures 4.7 and 4.6. In 4.7 restrict attention to the dashed lines which are for a bifunctional two-component system. The solid line shows results for mono-functional systems The upper left graph in both images shows the robustness property we discussed. More interesting is the upper right graph where it is shown that their model predicts that $\sigma^2_{R_P}$ is a concave function of I that reaches a global maximum. This is in direct contrast to what we found in Figure 4.6 where $\sigma^2_{R_P}$ increases in a linear fashion with L. We are inclined to hold more belief in the results of our model because it is a known rule that fluctuations scale with mean protein level [Bar-Even et al., 2006]. It is also reminiscent of the property of Poisson distributions: the mean is equal to the variance. Since this variance plays a role in the calculation of mutual information it is unsurprising that the two models also disagree on its dependence on L_{SS}. We predict an S-shaped curve for mutual information as a function of L at steady state. A puzzling fact is that the value in bits of mutual information is very low throughout the entire rather large range of L steady state values. It would be interesting to further investigate how much this can be altered by adjusting the model parameters. However, this is difficult as long as we do not understand the bifurcation behavior of the system. In the bottom-left graph the coefficient of

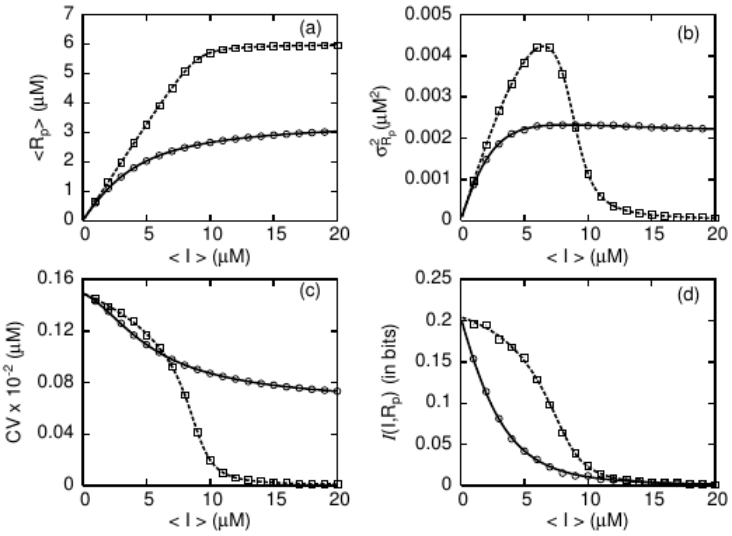


Figure 4.7: **Results from [Maity et al., 2014]** (a) Steady state R_p level, (b) steady state variance $\sigma^2_{R_P}$, (c) steady state squared coefficient of variation $\sigma^2_{R_P}/\langle R_P \rangle$ and (d) steady state mutual information $I(I, R_p)$ as a function of mean extra-cellular inducer level. In all panels, solid (with open circles) and dashed (with open squares) lines are for monofunctional and bifunctional system, respectively. The symbols are generated using stochastic simulation algorithm and the lines are due to theoretical calculation.

variation (σ/μ) is plotted. This provides a nice estimate for the size of fluctuations compared to the mean protein level.
It is likely that the difference in results stems from the fact that [Maity et al., 2014] exclude complex formation. However, to be sure two steps remain to be taken. First, Gillespie simulations should be run on the more detailed system considered here. If this returns similar results as in Figure 4.6, then it is unlikely that calculation error is the source of the different results. Second, it should be investigated if the approach used in this text when applied to the system without complex formation returns the same results as in [Maity et al., 2014] and Figure 4.7.

5 | Considerations about noise and errors

In the previous chapter we discussed the correlation between a signal L and the output of a TCS R_P through the linear noise approximation. However, we must not lose sight of the fact that a cell "makes decisions" based on the varying concentration of R_P, think of gene expression. We performed our calculations using mutual information, which gives a sense of the quality of the signal transduction process and could be interpreted as the logarithm of the number of states that can be distinguished in the signal. What it does not tell us is the likelihood of a single cell having a correct internal concentration - by correct we mean according to the deterministic input-output relationship. This is of particular biological importance though, since an experimental biologist may want to predict how many of an ensemble of cells make correct decisions.

This chapter is a short exploratory chapter in contrast to the previous ones, concerned with questions that mutual information is not really equipped to answer but that are relevant to biology. Specifically, we would like to answer the questions: based on the stochastic relationship between L and R_p, how often does a cell have the correct internal representation of the current level of L, and importantly, what does right mean in this context? In other words, what can we say about the probability that the cell infers the right concentration L based on $R_P(L)$? Additionaly, we would like to show that mutual information gives faulty answers to such questions. Put differently, can we devise two sets of noise/input distributions such that the mutual information in both channels is equal but the probability of faulty inference is not equal? If we can find such a situation then it definitively shows that mutual information in a way obscures relevant information for biologists and is not equipped to this task.

Since this is a somewhat different way of approaching the subject than is usual in the literature we must first establish a framework for thinking about these questions, starting out with as few assumptions as possible and expanding the model step by step. We begin by simply assuming that there is a deterministic communication channel between L and R_P and the only thing we take into account is that R_P is linearly related to L: $R_P = gL$ (the robustness property). Under this model, an input always leads to the same output and there is no possibility of error in the internal representation. Errors can arise in two ways: through the addition of noise to the communication channel or through variability in L itself.

5.1 Deterministic inputs and noise

First we consider a model of deterministic inputs but with noise in the channel. This model is a general form of the Gaussian channel we saw before:

$$R_P = gL + \epsilon,$$

where L is deterministic and ϵ is a random variable with some pdf $f_\epsilon(e)$. How do we define errors in this model? R_P has a distribution with mean $gL + \langle\epsilon\rangle$ and variance equal to that of ϵ. The best internal representation would be to have $R_P = gL$. In the case of a continuous noise variable ϵ, the probability of measuring that specific value is zero, so we can only speak of ranges of R_P being correct or incorrect.
In terms of R_P we can define an error by calculating the probability of being farther than δ of the optimal value gL. This is doable by simply integrating the pdf of the noise

$$\mathbb{P}(\text{error}) = 1 - \int_{-\delta}^{\delta} f_\epsilon(e)de.$$

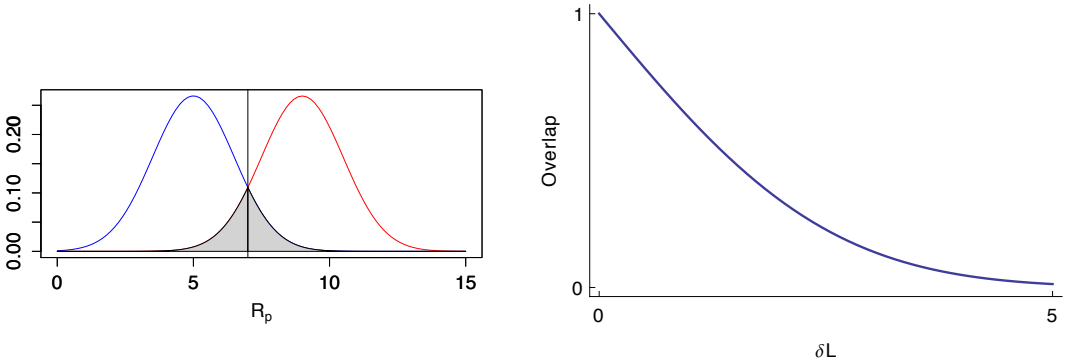
Note that we could expand this concept by adding weight factors which scale with the distance from the optimal value.

5.1.1 Two deterministic inputs and noise

Continuing, we consider two deterministic input levels L_1 and L_2, both leading to two distinct output distributions (when $L_1 \neq L_2$). Note that the only difference between these distributions will be their mean. We can imagine these inputs to be successive in time or parallel experiments. How do we define an error in this model? We could do the same as in the last model and for each input separately integrate over the probability of measuring at least δ away from the optimal values. But this would not be very interesting since it does not tell us how well the cell is able to discriminate between the two inputs.
New in this model is the possibility for overlap between the two distributions. What does this overlap signify? The overlapping probability mass gives the probability that the two different inputs lead to an output in the same range. If the overlapping probability mass is p then for that fraction of measurements we will not be able to tell with certainty which input was given. Therefore this area could be a good measure for the uncertainty a cell has about which input was given. There need not be any overlap between two R_P distributions. This situation can be created by having a large gain g, by having a small noise variance, or by considering two input values that are far apart. In this case there is no chance of making a wrong inference.
There are several difficulties in trying to calculate the overlap between two distributions in general. However, under the current model the R_P distributions will have the same distribution with the same variance but with a different mean, which makes things easier. The easiest distribution to find analytic results for is the Gaussian distribution. Again, it is a reasonable assumption for noise to be Gaussian either for fundamental reasons or through application of the central limit theorem.

5.1.2 Overlap between two Gaussian densities

Consider the two Gaussian distributions $N(gL, \sigma^2_{R_P})$ and $N(g(L+\delta L), \sigma^2_{R_P})$, which we assume to have the same variance. The difference in the mean level of R_P between these distributions is $\delta R_P = g\delta L$. The intersection of these Gaussian distributions is at $R_P = gL + \frac{g\delta L}{2}$ the average of the two means; see Figure 5.1a.



(a) (b)

Figure 5.1: (a) **Overlap between Gaussian densities**. Illustration of making a wrong inference using Gaussian distributions for the concentration of R_P. Blue is the distribution for low L and red is the distribution for high L concentration. They gray area corresponds to a wrong inference as explained in the main text. (b) **Overlap as a function of δL.** Plot of equation (5.1) for $g = 1$ and $\sigma^2_{R_P} = 1$.

The overlap between these two distributions equals twice the probability that the distribution centered at gL is larger than this intersection point. Or, in symbols

$$O(\delta L, g, \sigma^2_{R_P}) = 2\int_{gL+\frac{g\delta L}{2}}^{\infty} \Phi(r)dr = 2\left[1 - \Psi\left(gL + \frac{g\delta L}{2}\right)\right].$$

Here, we use $\Phi(r)$ to denote the pdf of the normal distribution with mean gL and variance $\sigma^2_{R_P}$, $\Psi(r)$ to denote its CDF and $O(\delta R_P, g, \sigma^2_{R_P})$ to denote the overlapping probability mass for a given mean distance δR_P, g and variance. We can rewrite this result in terms of the error function (A.13). Since in general $\Psi(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)\right]$ we see that

$$\begin{aligned} O(\delta L, g, \sigma^2_{R_P}) &= 2\left[1 - \frac{1}{2}\left[1 + \text{erf}\left(\frac{gL + \frac{g\delta L}{2} - gL}{\sqrt{2\sigma^2_{R_P}}}\right)\right]\right] \\ &= 1 - \text{erf}\left(\frac{g\delta L/2}{\sqrt{2\sigma^2_{R_P}}}\right). \end{aligned} \tag{5.1}$$

This function is plotted in Figure 5.1b as a function of δL. As is intuitively obvious, when $\delta L = 0$ the overlap is equal to 1 and the larger the distance between the two means the smaller the overlap becomes.

5.2 Stochastic input with noise

We will now consider stochastic input values so that $R_P = gL + \epsilon$ where both L and ϵ are random variables with densities $f_L(x)$ and $f_\epsilon(e)$. In the deterministic input case we could ask: given two input values of L, what is the overlap between the two output distributions in R_P? We can ask

the same question here but we now also know something about how likely certain input values are to occur since L has a probability distribution. Assuming L has a continuous distribution, we cannot say anything about specific values. However, for the Gaussian distribution we can say something about the difference δL between two successive draws. We can calculate the overlap between two output distributions based on how likely two successive *uncorrelated* draws from a Gaussian distributions are to be δL apart. This will result in an expected overlap in the output distributions for two such uncorrelated draws.

5.2.1 Expected overlap between two Gaussian densities originating from a Gaussian signal

If we assume L and ϵ to have Gaussian distributions, then R_P also has a Gaussian distribution. In (5.1) we saw that the overlap between two distributions of R_P based on two input values of L depends on the difference in L values, δL, g, and the standard deviation of R_P, σ_{R_P}. To determine the *average overlapping probability mass* over all possible values of δL, we first have to determine the probability distribution for δL and use these to weigh the overlap and integrate over δL. What is the probability distribution of δL? Phrased differently, we want to know the probability distribution for the difference of two Normal distributions. This has quite a simple answer; when X and Y are two normally distributed random variables with (possibly) different means and variances, this distribution is the so-called Normal difference distribution [1]:

$$f_{X-Y}(u) = \frac{\exp\left[-(u-\mu_X+\mu_Y)^2/[2(\sigma_X^2+\sigma_Y^2)]\right]}{\sqrt{2\pi(\sigma_X^2+\sigma_Y^2)}}. \tag{5.2}$$

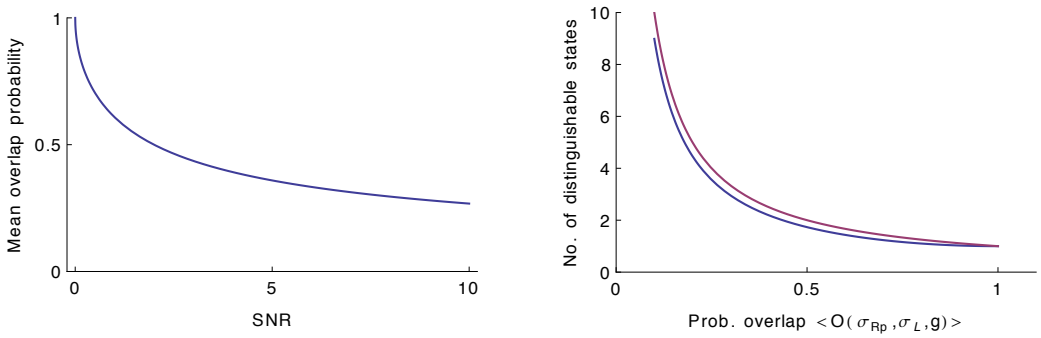
Important to note is that this result only holds if X and Y are independent. In essence this is just the well-known result that the sum (or difference) of two Gaussian distributions is a Gaussian distribution with the sum (or difference) of the means as mean and a variance that is the sum (in both cases) of the original variances.
In the case of δL we draw two samples L_1 and L_2 from the same distribution, so that the above result simplifies to

$$f_{L_1-L_2}(\delta L) = \frac{\exp\left[-\frac{(\delta L)^2}{4\sigma_L^2}\right]}{\sqrt{4\pi\sigma_L^2}}.$$

Observe that this is again a normal distribution with mean 0 and variance $2\sigma_L^2$ and this last equation is a function of δL.
We can now calculate the mean overlap over all values of δL by integrating [2]

$$\begin{aligned}\langle O(\sigma_{R_P}^2, \sigma_L^2, g)\rangle &= 2\int_0^\infty \left(1-\operatorname{erf}\left(\frac{g\delta_L/2}{\sqrt{2\sigma_{R_P}^2}}\right)\right)\frac{\exp\left[-\frac{\delta L^2}{4\sigma_L^2}\right]}{\sqrt{4\pi\sigma_L^2}}d\delta L \\ &= 1-\frac{2}{\pi}\text{Arctan}\left(\frac{g\sigma_L}{\sqrt{2\sigma_{R_P}^2}}\right) \qquad (5.3)\\ &= 1-\frac{2}{\pi}\text{Arctan}\left(\sqrt{\frac{\text{SNR}}{2}}\right). \qquad (5.4)\end{aligned}$$

[1] The entry on Wolfram MathWorld has an error in this formula, although it is correct in their Mathematica notebook!
[2] This argument is originally by F.J. Bruggeman



(a) (b)

Figure 5.2: (a) **The mean overlap probability**. mean overlapping probability mass (5.4) as a function of the SNR. (b) **The number of distinguishable states vs. the mean overlap probability**. In blue, (5.5) is plotted and in red $1/\langle O(\sigma^2_{R_P}, \sigma^2_L, g)\rangle$ is plotted.

The last equality follows from the definition of the signal-to-noise ratio SNR $= \frac{g^2\sigma^2_L}{\sigma^2_{R_P}}$. This is an intriguing result! In Figure 5.2a we plot this mean overlapping probability mass as a function of the SNR.
This graph agrees with our intuition. A SNR close to zero implies a relatively large amount of noise, meaning wide distributions of R_P which imply a large overlap. When the SNR is large, meaning there is relatively little noise, the R_P distributions are thin so there is little overlap.
We can go further than the result for the mean overlap distribution. Due to the simple expression for the mutual information in a Gaussian channel (2.28), we can also relate the MI, and thus the number of states of L that R_P can distinguish, to the mean overlapping probability mass. For mutual information measured in bits we have that

$$\begin{aligned}\text{No. Distinguishable states} &= 2^{I(R;S)} = 2^{\frac{1}{2}\log_2(1+SNR)} \\ &= \sqrt{1+SNR}.\end{aligned}$$

Rewriting (5.4) as SNR $= 2\tan^2\left(\frac{1}{2}\pi(\langle O(\sigma^2_{R_P}, \sigma^2_L, g)\rangle - 1)\right)$ and substituting the formula above, we see

$$\begin{aligned}\text{No. Distinguishable states} &= \sqrt{1+2\tan^2\left(\frac{1}{2}\pi(\langle O(\sigma^2_{R_P}, \sigma^2_L, g)\rangle - 1)\right)} \\ &= \sqrt{1+2\cot^2\left(\frac{\pi\langle O(\sigma^2_{R_P}, \sigma^2_L, g)\rangle}{2}\right)}.\end{aligned} \tag{5.5}$$

In Figure 5.2b, equation (5.5) is plotted against $\langle O(\sigma^2_{R_P}, \sigma^2_L, g)\rangle$ along with $1/\langle O(\sigma^2_{R_P}, \sigma^2_L, g)\rangle$. We see good agreement between the two graphs, indicating that $1/\langle O(\sigma^2_{R_P}, \sigma^2_L, g)\rangle$ gives a good estimate of the number of distinguishable states. However, it also tells us something about the probability of faulty inference which mutual information does not.
Several questions are still left unanswered here. First of all it is not entirely obvious why 1 divided by the overlap probability should be a good estimate of the number of distinguishable states. Second, it remains unclear why the graphs are not exactly the same. We seem to overestimate the number of states, i.e. the overlap is to small. This may have something to do with the fact that we use two inputs instead of more. Both these questions require further research.

5.3 Looking at errors in the L domain

A second way to look at errors is to consider what we can say about L given a measurement of $R_P = y$, i.e. what is the probability of inferring a certain range L is likely to be in based on observing R_P? We would like to say something about $\mathbb{P}(\alpha \leq L \leq \beta | R_p = y)$ and for this we need the conditional density $\mathbb{P}_L(x|R_P = y)$, which we can get by dividing the joint probability density by the marginal density of R_P:

$$f_L(x|R_P = y) = \frac{f_{L,R_P}(x,y)}{f_{R_P}(y)}$$

for all values of R_P such that $f_{R_P}(y) > 0$ and zero otherwise. Specifically, we can calculate the desired probability using

$$\mathbb{P}(\alpha \leq L \leq \beta | R_p = y) = \int_\alpha^\beta f_L(x|R_P = y)dx \tag{5.6}$$

$$= \frac{\int_\alpha^\beta f_{L,R_P}(x,y)dx}{\int_{-\infty}^{\infty} f_{L,R_P}(x,y)dx}. \tag{5.7}$$

The denominator in this equation is the total area under the joint probability density function at $R_P = y$. The numerator is the mass of that area where $\alpha \leq L \leq \beta$.
By choosing the right α and β, this calculation will give us the probability of inferring (approximately) the correct value of L given a measurement of intracellular R_P.
The main mathematical difficulties in this approach are (1) analytically finding the joint pdf and (2) analytically calculating this integral. Numerically however, given a standard noise distribution this calculation should be trivial.
A more fundamental problem with this approach is the choice of α and β. We would like these to be estimates of the range the signal can take in a certain state of the environment. For instance, if the environment can exist in a high and low sugar state and the low state means that the concentration falls between 5 and 10 mmol/L, then we would like to calculate $\mathbb{P}(5 \leq L \leq 10 | R_p = y)$. However, this requires knowledge of the extracellular concentration levels and it is difficult to imagine a cell actually implementing this calculation.

5.3.1 Mutual information and MMSE

Related to this inverse estimation problem introduced above, we found an interesting paper [Guo et al., 2005] by scouring the engineering literature. [Guo et al., 2005] find that mutual information and the MMSE (minimum mean squared error) in estimating the input given the output, satisfy a simple relationship regardless of the input distribution as long as they are related through additive Gaussian noise. This result holds in general as the next theorem shows, which we reproduce from [Guo et al., 2005] without proof.

Theorem 5.3.1. Let ξ (the noise) be standard Gaussian, independent of X. For every input distribution P_X (discrete or continuous) that satisfies $\mathbb{E}X^2 < \infty$,

$$\frac{d}{d\,\mathrm{SNR}} I(X;Y) = \frac{1}{2}\mathrm{MMSE}(X|Y). \tag{5.8}$$

Proof. See II-C in [Guo et al., 2005]. □

This result holds not only under arbitrary input signaling but also under the broadest setting of Gaussian channels, including discrete-time and continuous-time channels, and both scalar and

vector versions. First, note that both mutual information and the MMSE are increasing functions of the SNR. Thus, this relationship states that the rate at which mutual information increases as the SNR is increased is one-half of the MMSE value.
The error of an estimate, $f(Y)$, of X based on observing Y can be measured in mean-square sense (well-known from statistics):

$$\mathbb{E}\left[(X - f(Y))^2\right]. \tag{5.9}$$

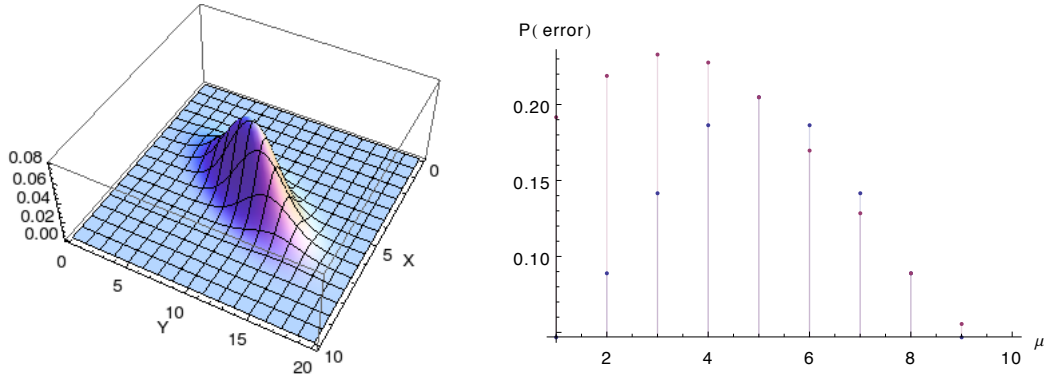
The minimum value of this estimator is referred to as the MMSE and is under weak regularity assumptions given by the so-called conditional mean estimator $\hat{X} = \mathbb{E}[X|Y]$ (See Lehmann and Casella, Corollary 4.1.2). As we know for the Gaussian channel with Gaussian input $f_X(x)$, we have that $I(X;Y) = \frac{1}{2}\log(1 + \text{SNR})$. Obviously $\frac{d}{d\,\text{SNR}} I(X;Y) = \frac{1}{2}\frac{1}{1+\text{SNR}}$. The MMSE for this case happens to be $\frac{1}{1+\text{SNR}}$ thus $\frac{d}{d\,\text{SNR}} I(X;Y) = \frac{1}{2}\text{MMSE}(X|Y)$.
Note that this relationship is intuitively correct in the sense that for SNR $\to \infty$ the left side goes to zero in the standard Gaussian channel we have been considering, since $\frac{d}{d\,\text{SNR}}\frac{1}{2}\log(1 + \text{SNR}) = \frac{1}{2}\frac{1}{1+\text{SNR}}$. Therefore, the estimation error also goes down to zero. This makes sense since maximizing SNR is equivalent to maximizing the mutual information which should reduce uncertainty about the input per its definition in terms of a difference of entropy.
The result seems to suggest a relationship between mutual information and this reverse inference problem of estimating the input given the output. What remains unclear, is how we can specifically use this to say something useful about the process of signal transduction. The fact that relationships like these exist in the engineering literature yet have not appeared in the biological literature (as far as we can tell) suggests that there may be more potential gems hidden in the engineering literature that remain to be discovered.

5.4 Mutual information and error probability in gene expression

Here, we consider error probability in relation to a simplistic view of gene expression.

5.4.1 Switch-like gene expression in a Gaussian channel

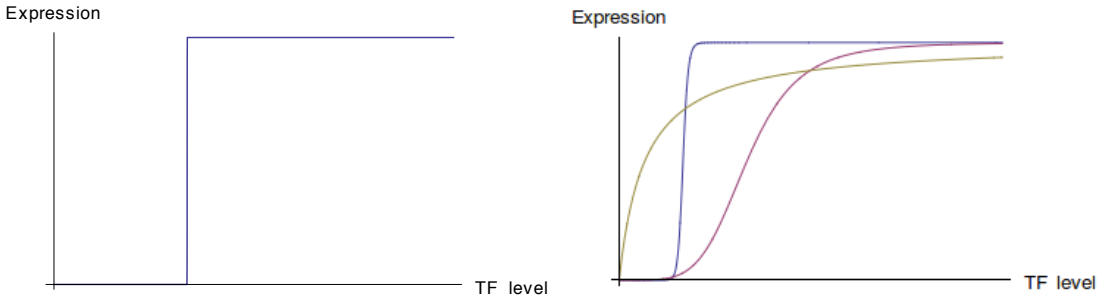
Consider once again the Gaussian channel setting and a hugely simplified model of gene expression where a gene is expressed if a certain level of Y is exceeded $Y > \bar{y}$. Consider further, as done before, that this Y variable is a cell's internal representation of an extracellular signal X. Ideally, given a level of X considered high, $\bar{x}$, Y should be high as well, so exceeding $\bar{y}$. If this is not the case, we qualify this as an error since the cell would respond as if the signal is high (low) even though in reality the signal was low (high).
By keeping the signal-to-noise ratio constant the mutual information remains constant. But, as shown below, changing the variances while keeping SNR constant or changing the mean of the input distribution changes the probability of making an error under this model. This indicates that under this highly simplified model of gene expression mutual information does not track error probability. This is intuitively obvious, since, as noted before, mutual information ignores the mean whereas for gene expression the mean copy number of a transcription factor clearly matters!
Under the Gaussian channel assumption, the X and the noise both have Gaussian densities and since they are independent we can calculate the joint distribution between X and Y using $f_{X,Y}(x,y) = f_{Y|X}(y|x)f_X(x)$. Under this simplified model, we simply need to calculate the



(a) (b)

Figure 5.3: (a) **Joint distribution of two Gaussians**. A 3D plot of the joint distribution of two Gaussian random variables X and Y with $\mu_X = 5$, $\sigma = 1$, $\sigma_Y = 2$, $g = 2$. (b) **The error probability for changing values of** μ_X. In blue: $\sigma_N = 3$, $g = 2$, $\sigma_X = 2$. In red: $\sigma_N = 4.5$, $g = 2$, $\sigma_X = 3$. In both cases SNR $= 1.5$.

amount of probability mass of this joint density function that falls in the wrong two quadrants (high x and low y and vice versa). In Figure 5.3, we show the joint distribution of X and Y and a plot of the changing error probability for varying mean levels of X and two different settings for the variances while keeping the signal-to-noise ratio constant.



(a) (b)

Figure 5.4: (a) **Switch-like gene expression**. At a certain transcription factor level gene expression switches on. This is equivalent to taking $n \to \infty$ in the Hill activation function. (b) **Smooth gene expression for various values of** n. The smaller n becomes the less steep the activation function is.

We could expand this model to incorporate smooth gene expression of the form through something like a Hill activation function

$$\text{Expression} = \frac{y^n}{K^n + y^n}$$

for some specific value of n which dictates the steepness of activation; see Figure 5.4. However, the argument stays the same and shows that mutual information simply is not equipped to answer questions like this.

6 | Conclusion

In this thesis, we concerned ourselves with the topic of information processing in two-component signal transduction systems. Here we briefly review the results we found and what avenues remain for future research.

6.1 Summary of results and future work

Results and future work have been divided into the separate research questions we looked at in this text.

How can we quantify the knowledge or information a cell has about its environment? And: What is a good measure of the quality of this signal transduction process? We considered this in Chapter 2. We saw how we can model the process of signal transduction through a communication channel between two random variables that have some dependence on each other; see Figure 2.1. The noise in such a channel reduces a cell's ability to discern different input levels. Since we want to somehow quantify the quality of this signal transduction process, we turned to mutual information. Mutual information says something about the reduction in uncertainty (entropy) about the input given the output (and vice versa because of the symmetry property) and is therefore a good measure of the quality of communication. In addition, we showed its various nice mathematical properties including the recently published result of self-equitability. Mutual information is justifiably used as a quality measure of communication channels and will most likely be around for quite some time.

How does mutual information relate to the probability of the cell's internal representation being wrong? We considered this in Chapter 5. Closely related to the process of information transmission and signal transduction is decision-making and gene expression since cells cannot make appropriate decisions without exploiting information from external stimuli. A key shortcoming of mutual information is that it does not explicitly tell us how likely it is that a cell makes a right inference. We considered various ways of looking at this question by looking at overlap distributions in signal system output (R_P) and by thinking about the inverse estimation procedure of saying something about the $f_{X|Y}(x, y)$ distribution.
In the context of Gaussian overlap distributions we found a surprising result (5.4) that seems to be a good estimate of the number of distinguishable states in the input 5.2b. The reasons for this remain unclear and it remains an avenue for further research.
In the context of looking at the probability $\mathbb{P}(\alpha \leq L \leq \beta | R_p = y)$ we found a relation that links mutual information to the minimal mean squared error (MMSE) in estimating the input given the output [Guo et al., 2005]. We do not yet fully understand how to apply this to signal

transduction, however the fact that this relationship was found in an engineering paper suggests there may be more to discover that has not yet been applied to biology.
Lastly, we considered mutual information in relation to error probability in simplified gene expression. The point of this was mainly to show that there are interesting questions in signal transduction and gene expression for which mutual information is not the right tool. We showed this by assuming switch-like gene expression, but the argument can be extended to smooth expression functions. The probability mass falling in the wrong two quadrants can be changed while keeping mutual information in a Gaussian channel; see Figure 5.3.

Can we analytically and numerically calculate linear noise approximation for a specific two-component system? In Chapter 3 we considered stochastic kinetics. We introduced master equations and Fokker-Planck equations including their stationary solution. We then discussed the proof and implementation of the system size expansion, as originally done by [Van Kampen, 2007] and [Elf and Ehrenberg, 2003].
In Chapter 4 we then considered a specific bifunctional two-component system of medium complexity. We applied the linear noise approximation to two variants of this system: the constant L model and the model including synthesis and degradation of L.
In [Maity et al., 2014], a TCS model is considered without taking into account any complexes. We considered a model that takes into account more of the molecular detail, although not all; see Figure 4.3a. We specifically choose the bifunctional version of the TCS because it has the nice robustness property, 1 less reactant and beneficial results relating to the amount of MI it can handle (see [Maity et al., 2014].
Numerically it is quite easy to perform the linear noise approximation on this system and find the covariance matrix. Analytically we used an LU decompositon on A and solved in two steps. The analytical result for the variance of R_P we found was a massive symbolic result and thus for the dynamic L model we abandoned the analytical approach and focused on numerical calculations.

How does the output variance change with model parameters? In Figure 4.4 we plotted the dependecies of $\mathrm{Var}(R_P)$ in terms of all model parameters for the constant L model. In Figure 4.5 we do the same for the model with L dynamics included. Interesting to note is that in the dynamic L model the variance in R_P rises approximately linearly with L at steady state 4.6. In the constant L model this relationship is exactly linear. In both models it can also be seen that parameters belonging to the same reaction increase and decrease the variance in opposing pairs. In addition, reaction 3 seems to have no effect on the variance in both models.
For the dynamic L model there is an open question relating to the model's bifurcation behaviour. We do not yet understand how to guarantee that a certain parameter set leads to the non-trivial steady state.

What consequences does this have for the mutual information between the input and output? We can calculate the mutual information between L and R_P for specific parameter sets if we assume a Gaussian channel setting and using the $I(X;Y) = -\frac{1}{2}\log_2\left(1-\rho^2\right)$ result. For the dynamic L model, we predict that mutual information as a function of L at steady state has an S shaped form; see Figure 4.6. A good intuitive explanation for this is lacking as of yet.
For the constant L model we did not perform this calculation since we would have to assume a variance for L which seems quite arbitrary. Since the other results seem to carry over quite well between the two models, we would expect this to return similar results to Figure 4.6.

How do our results compare with the recent publication [Maity et al., 2014]? In [Maity et al., 2014] a less detailed two-component system is considered that does not take into account complex formation. They find that the variance in R_P increases with L, reaches an optimum and

goes back down. We do not find this for the system with constant L nor for the system with dynamic L. To facilitate easy comparison, we reproduced the 4 panel plot from their paper in Figure 4.6. We reproduced their results in Figure 4.7.
It is likely that the difference lies in the fact that [Maity et al., 2014] exclude complex formation. However, two steps remain to be taken to be certain. First, Gillespie simulations should be run on the more detailed system considered here. If this returns similar results as in 4.6 then it is unlikely that calculation error is the source of the different results. Second, it should be investigated if the approach used in this text when applied to the system without complex formation returns the same results as in [Maity et al., 2014].

A | Appendix

A.1 The correlation coefficient

Pearson's product moment correlation coefficient measures the existence of a *linear* relationship between two random variables. It takes on values in the range $[-1, 1]$ based on it being a negative (downward slope) linear relation or a positive (upward slope) relation.

Definition A.1.1. If X and Y are random variables with variances σ_X^2 and σ_Y^2 and covariance $\mathrm{Cov}(X, Y) = \sigma_{XY}^2$, then the *correlation coefficient* of X and Y is

$$\rho = \frac{\sigma_{XY}^2}{\sigma_X \sigma_Y}. \tag{A.1}$$

If $\rho = 0$ we say that X and Y are uncorrelated.

Theorem A.1.1.

$$-1 \leq \rho \leq 1$$

Proof. Let $Z = \frac{Y}{\sigma_Y} - \rho\frac{X}{\sigma_X}$. Then

$$\begin{aligned}
\mathrm{Var}(Z) &= \mathrm{Var}\left(\frac{Y}{\sigma_Y}\right) + \mathrm{Var}\left(-\rho\frac{X}{\sigma_X}\right) + 2\,\mathrm{Cov}\left(\frac{Y}{\sigma_Y}, -\rho\frac{X}{\sigma_X}\right) \\
&= \left(\frac{1}{\sigma_Y}\right)^2 \sigma_Y^2 + \rho^2 \left(\frac{1}{\sigma_X}\right)^2 \sigma_X^2 - 2\rho\frac{\sigma_{XY}}{\sigma_X \sigma_Y} \\
&= 1 + \rho^2 - 2\rho^2 \\
&= 1 - \rho^2 \geq 0.
\end{aligned}$$

The last inequality shows the desired result. □

One can show that $\rho = \pm 1$ if and only if $Y = aX + b$ with probability 1 for $a \neq 0$ and any b; see any probability textbook.

A.2 The signal-to-noise ratio vs. the correlation coefficient

There exists an interesting relationship between the SNR and the correlation coefficient (ρ) that will allow us to rewrite (2.28) in terms of ρ. In this section we derive this relationship in two ways.

A.2.1 Geometric argument

We can rewrite the Gaussian channel assumption by assuming two separate gain parameters a and b for the signal and the noise as:

$$Y = aX + b\xi.$$

Here we are assuming that X, Y and ξ have been normalized to have zero mean and unit variance through their Z-scores. Now note that

$$\begin{aligned}\sigma_R^2 &= a^2\sigma_S^2 + b^2\sigma_\xi^2 \\ 1 &= a^2 + b^2.\end{aligned}$$

Denoting a^2 by α we see that

$$\text{SNR} = \frac{\alpha\sigma_S^2}{(1-\alpha)\sigma_\xi^2} = \frac{\alpha}{1-\alpha} \tag{A.2}$$

The correlation coefficient can be defined as the cosine of the angle θ between the two vectors of samples drawn from the two random variables

$$\rho = \cos(\theta) = \frac{x \cdot y}{||x||||y||}.$$

Since we assumed that the signal and the noise are independent, we can see them as orthogonal basis vectors for the response. Thus we can write

$$\tan(\theta) = \frac{b}{a} = \frac{\sqrt{1-\alpha}}{\alpha}.$$

Rewriting

$$\begin{aligned}\tan(\theta) &= \frac{\sin(\theta)}{\cos(\theta)} = \frac{\sqrt{1-\cos^2(\theta)}}{\cos(\theta)} \\ &= \frac{\sqrt{1-\rho^2}}{\rho}\end{aligned}$$

we have that $\alpha = \rho^2$. Thus we can now express the signal-to-noise ratio in terms of the correlation coefficient

$$\text{SNR} = \frac{\rho^2}{1-\rho^2}.$$

We can therefore also express the mutual information in terms of ρ

$$\begin{aligned}I(X;Y) &= \frac{1}{2}\log_2\left(1 + \frac{\rho^2}{1-\rho^2}\right) \\ &= -\frac{1}{2}\log_2\left(1-\rho^2\right)\end{aligned} \tag{A.3}$$

A.2.2 Entropy argument

A second way to derive this relationship is by considering the joint entropy of X and Y when both are Gaussian. In this case simply substituting the Gaussian distribution into the continuous version of (2.8) reveals that

$$h(X;Y) = 1 + \log(2\pi) + \frac{1}{2}\log(|\det \mathbf{\Sigma}|) \tag{A.4}$$

where $\mathbf{\Sigma}$ is the covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix} \tag{A.5}$$

so that $|\det \mathbf{\Sigma}| = \sigma_X^2\sigma_Y^2(1-\rho^2)$. Remembering (2.2.1) we see that

$$I(X;Y) = h(X) + h(Y) - h(X;Y) \tag{A.6}$$

$$= -\frac{1}{2}\log\left(1-\rho^2\right) \tag{A.7}$$

A.3 The Gershgorin circle theorem

First published in 1931 by the Russian mathematician Gershgorin, this theorem can be used to bound eigenvalues of a square (complex) matrix .

Theorem A.3.1. Let A be a (complex) n x n matrix, with entries a_{ij}. For $i \in \{1,\dots,n\}$ let $R_i = \sum_{j\neq i}|a_{ij}|$ be the sum of the absolute values of the non-diagonal entries in row i. Let $D(a_{ii}, R_i)$ be the closed disc centered at a_{ii} with radius R_i. Such a disc is called a Gershgorin disc and every eigenvalue lies in at least one such disc.

Proof. Let λ be an eigenvalue of A and x its corresponding eigenvector. Choose i such that $|x_i| = \max_j|x_j|$. Since x cannot be 0, $|x_i| > 0$. Now $Ax = \lambda x$, or looking at the i-th component $(\lambda - a_{ii})x_i = \sum_{j\neq i} a_{ij}x_j$. Taking the norm on both sides gives $|\lambda - a_{ii}| = |\sum_{j\neq i}\frac{a_{ij}x_j}{x_i}| \leq \sum_{j\neq i}|a_{ij}|$. □

Relevant for the discussion in the main text is the corollary that when applied to A^T, this result holds for the columns of A as well. This is easily seen because, as the identity matrix is symmetric we have that

$$\det(A^T - \lambda I) = \det\left((A-\lambda I)^T\right) = \det(A-\lambda I)$$

since $\det A = \det A^T$. Therefore A and A^T have the same characteristic polynomial and the same eigenvalues.

A.4 LU decomposition of a tridiagonal matrix

LU decomposition factors a matrix as the product of a lower triangular matrix L and an upper triangular matrix U. For tridiagonal matrices, there exists an especially simple algorithm for LU-decomposition. Consider the following notation for the entries in these three matrices

$$A = \begin{pmatrix} b_1 & c_1 & & 0 & \\ a_2 & b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{n-1} & b_{n-1} & c_{n-1} \\ & 0 & & a_n & b_n \end{pmatrix} \tag{A.8}$$

$$LU = \begin{pmatrix} 1 & & & 0 & \\ l_2 & 1 & & & \\ & \ddots & \ddots & \ddots & \\ & & l_{n-1} & 1 & \\ & 0 & & l_n & 1 \end{pmatrix} \begin{pmatrix} u_1 & c_1 & & 0 & \\ & u_2 & c_2 & & \\ & & \ddots & \ddots & \\ & & & u_{n-1} & c_{n-1} \\ & 0 & & & u_n \end{pmatrix}. \tag{A.9}$$

By simply performing the matrix multiplication on the right-hand side and equating entry by entry left and right we see that the following recursive relationships hold for l_k and u_k:

$$a_k = l_k u_{k-1} \qquad \Longrightarrow \qquad l_k = \frac{a_k}{u_{k-1}}, \qquad k \in \{2, \ldots, n\} \tag{A.10}$$

$$b_k = l_k c_{k-1} + u_k \qquad \Longrightarrow \qquad u_k = b_k - l_k c_{k-1}, \qquad k \in \{1, \ldots, n\}. \tag{A.11}$$

Where we use the convention that $l_1 = 0$ and $c_0 = 0$.

A.5 The error function erf(x)

The (Gauss) error function comes up from time to time in statistics and probability and may be interpreted as the probability of a random variable with a Gaussian distribution of mean 0 and variance $\frac{1}{2}$ falling in the range $[-x, x]$,

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \, \mathrm{d}t \tag{A.12}$$

$$= \frac{1}{\sqrt{\pi}} \int_{-x}^x e^{-t^2} \, \mathrm{d}t. \tag{A.13}$$

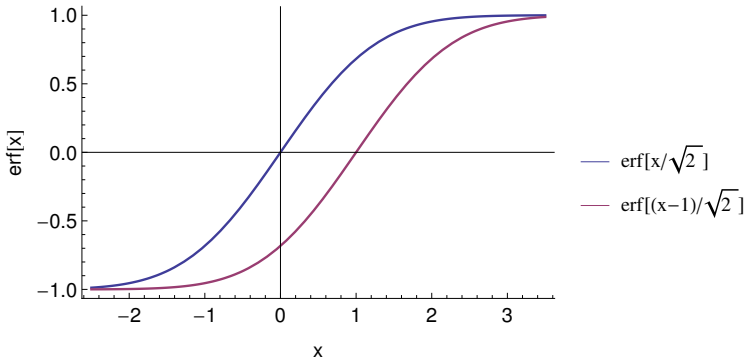


Figure A.1: **The error function**. Plot of $\operatorname{erf}(x/\sqrt{2})$ and $\operatorname{erf}((x-1)/\sqrt{2})$.

To see the relationship between the CDF of a normal distribution and $\operatorname{erf}(x)$ consider the CDF of a random variable X with a standard normal distribution

$$\begin{aligned}
\Psi_X(x) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{t^2}{2}\right) dt \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^0 \exp\left(-\frac{t^2}{2}\right) dt + \frac{1}{\sqrt{2\pi}} \int_0^x \exp\left(-\frac{t^2}{2}\right) dt \\
&= \frac{1}{2} + \frac{1}{2} \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right).
\end{aligned}$$

For non-standard normal distributions, we have that $\Psi_{\frac{X-\mu}{\sigma}}(x) = \Psi_X(\frac{x-\mu}{\sigma})$ and therefore

$$\Psi_{\frac{X-\mu}{\sigma}}(x) = \frac{1}{2} + \frac{1}{2} \operatorname{erf}(\frac{x-\mu}{\sqrt{2\sigma^2}}). \tag{A.14}$$

References

[Attneave, 1954] Attneave, F. (1954). Some informational aspects of visual perception. *Psychological review*, 61(3):183.

[Bar-Even et al., 2006] Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'Shea, E., Pilpel, Y., and Barkai, N. (2006). Noise in protein expression scales with natural protein abundance. *Nature Genetics*, 38(6):636–643.

[Bartels and Stewart, 1972] Bartels, R. H. and Stewart, G. W. (1972). Solution of the matrix equation AX+ XB= c [f4]. *Communications of the ACM*, 15(9):820–826.

[Cover and Thomas, 1991] Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. Wiley-Interscience.

[Elf and Ehrenberg, 2003] Elf, J. and Ehrenberg, M. (2003). Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome Research*, 13(11):2475–2484.

[Gardiner, 1985] Gardiner, C. W. (1985). *Handbook of stochastic methods*, volume 3. Springer Berlin.

[Gillespie, 1977] Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361.

[Goel and Richter-Dyn, 1974] Goel, N. S. and Richter-Dyn, N. (1974). *Stochastic models in biology.*

[Guo et al., 2005] Guo, D., Shamai, S., and Verdú, S. (2005). Mutual information and minimum mean-square error in gaussian channels. *Information Theory, IEEE Transactions on*, 51(4):1261–1282.

[Haykin, 1994] Haykin, S. (1994). *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

[Hayot and Jayaprakash, 2004] Hayot, F. and Jayaprakash, C. (2004). The linear noise approximation for molecular fluctuations within cells. *Physical biology*, 1(3-4):205–210.

[Heinrich and Schuster, 1996] Heinrich, R. and Schuster, S. (1996). *The Regulation of Cellular Systems*. Springer.

[Horn, 1994] Horn, R. A. (1994). *Topics in Matrix Analysis*. Cambridge University Press.

[Igoshin et al., 2008] Igoshin, O. A., Alves, R., and Savageau, M. A. (2008). Hysteretic and graded responses in bacterial two-component signal transduction. *Molecular Microbiology*, 68(5):1196–1215.

[Kinney and Atwal, 2014] Kinney, J. B. and Atwal, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, page 201309933.

[Linsker, 1988] Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, 21(3):105–117.

[Maarleveld et al., 2013] Maarleveld, T. R., Olivier, B. G., and Bruggeman, F. J. (2013). StochPy: A comprehensive, user-friendly tool for simulating stochastic biological processes. *PLoS ONE*, 8(11):e79345.

[Maity et al., 2014] Maity, A. K., Bandyopadhyay, A., Chaudhury, P., and Banik, S. K. (2014). Role of functionality in two-component signal transduction: A stochastic study. *Physical Review E*, 89(3):032713.

[Pahle et al., 2012] Pahle, J., Challenger, J. D., Mendes, P., and McKane, A. J. (2012). Biochemical fluctuations, optimisation and the linear noise approximation. *BMC Systems Biology*, 6(1):86.

[Perkins and Swain, 2009] Perkins, T. J. and Swain, P. S. (2009). Strategies for cellular decision-making. *Molecular systems biology*, 5:326.

[Reshef et al., 2013] Reshef, D., Reshef, Y., Mitzenmacher, M., and Sabeti, P. (2013). Equitability analysis of the maximal information coefficient, with comparisons. *arXiv preprint arXiv:1301.6314*.

[Rhee et al., 2012] Rhee, A., Cheong, R., and Levchenko, A. (2012). The application of information theory to biochemical signaling systems. *Physical Biology*, 9(4):045011.

[Rieke, 1999] Rieke, F. (1999). *Spikes: exploring the neural code*. MIT press.

[Shannon, 1948] Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.

[Shinar et al., 2007] Shinar, G., Milo, R., Martínez, M. R., and Alon, U. (2007). Input–output robustness in simple bacterial signaling systems. *Proceedings of the National Academy of Sciences*, 104(50):19931–19935.

[Stock et al., 2000] Stock, A. M., Robinson, V. L., and Goudreau, P. N. (2000). Two-component signal transduction. *Annual review of biochemistry*, 69:183–215.

[Thomas et al., 2013] Thomas, P., Matuschek, H., and Grima, R. (2013). How reliable is the linear noise approximation of gene regulatory networks? *BMC Genomics*, 14(Suppl 4):S5.

[Tkacik et al., 2008] Tkacik, G., Callan, Jr, C. G., and Bialek, W. (2008). Information capacity of genetic regulatory elements. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 78(1 Pt 1):011910.

[Van Kampen, 2007] Van Kampen, N. (2007). *Stochastic Processes in Physics and Chemistry, Third Edition (North-Holland Personal Library)*. North Holland.

[Walczak and Tkačik, 2011] Walczak, A. M. and Tkačik, G. (2011). Information transmission in genetic regulatory networks: a review. *Journal of Physics: Condensed Matter*, 23(15):153102. arXiv:1101.4240 [physics, q-bio].

[Yeung, 2010] Yeung, R. W. (2010). *Information Theory and Network Coding*. Springer, New York, softcover reprint of hardcover 1st ed. 2008 edition.