CHAPTER 1

INTRODUCTION

1.1 Cellular communication

The basic building block of all living organisms is the cell. Indeed, a single cell can be a living organism by itself, like the unicellular bacteria from the kingdom of the prokaryotes. Next to these unicellular organisms, multicellular organisms exist. In these organisms multiple cells group, act and communicate in order for the organism to survive. The realization that cells are the fundamental units dates back to the German biologists Theodor Schwann and Matthias Schleiden, who in 1839, formulated a theory which posits that the cell, both in animals and plants, is the fundamental unit of life [1].

Since then more than a 170 years have passed in which our knowledge of these cells has increased dramatically. One important realization is that cells, although being coined the building block of life, themselves consist of many other smaller structures. The main difference between these structures — e.g. membranes, the nucleus, DNA, motors — and the cell is that a cell is capable of reproduction, while the individual building blocks themselves are not. A cell therefore is alive, while its building blocks are lifeless — a fascinating observation on its own. The building blocks themselves consist of molecules like proteins, lipids, receptors and amino-acids, which are the essential lego-parts of all cellular structures and thus the cell.

These smaller building cellular structures are essential for any cell to survive. If we focus on a single cell, for example a bacterium, we observe that it has to overcome many challenges in its struggle for survival. Only looking at the process of cell-division, it is directly clear that the cells faces challenges such as: When does the cell “know” that division should take place? How does the cell spatially divide itself in equal halves, and as important, ensure that essential proteins and organelles are equally distributed over each half? Where and why do structures form that were not present before, like the FtsZ-ring in bacterial cell-division?

Not only for cell division these kinds of questions arise. Bacteria (e.g. E. coli) are capable of detecting sugar concentrations in their environment and even move to locations with a larger abundance of sugar, a process called chemotaxis [2]. This is a remarkable achievement,
recognizing the limited size of the bacterium. The same bacterium is capable of surviving on many different sugar sources [3], including glucose and lactose, but in general only the enzyme for the conversion of glucose to energy (ATP) is present. This means that if only lactose is available, the bacterium has to change its internal metabolism path, e.g. produce new enzymes that can act on lactose. Another example of sensing is provided by yeast (S. cerevisiae) $\alpha$-cells, which secretes small pheromone molecules ($\alpha$-factor), to mark their presence to yeast $\alpha$-cells, which eventually leads to mating of these two cells [4].

Clearly many processes continuously take place in cells, a small list based upon the above examples could include: the sensing of the environment by membrane proteins (receptors), initiation or termination of gene expression by proteins (transcription factors), activation of motors (myosin, kinesin), or the changes in the cellular structure by protein (actin, dynein). All of these processes are vital for the cell in its order to survive.

Many, if not all, of these processes require some form of communication or signal transfer between the cell and its environment, or between processes entirely within the cell. Considering the above examples, here sugar levels lead to cell motility, sensing $\alpha$ or $\alpha$ factor leads to mating, and the start of division leads to an internal re-organization of cellular components. Cells have developed special networks that essentially have this task: transmitting signals from the extracellular to the intracellular environment or transmitting signals within the cell. These networks are referred to as signaling cascades (or signal transduction cascades) and these cascades are the main topic of this thesis.

1.1.1 Signaling cascades

The goal of a signaling cascade is to transmit a signal, through multiple intermediate steps to a response. The concept of passing a signal around through multiple steps leading to a response is very well mimicked in a very famous children’s game: the telephone game.

In the telephone game a group of children sit in half a circle. At one end, a child gets a message — in general a complicated or hilarious sentence — and has to whisper this message to his or her neighbor, but such that the other children can not hear it. This process continues until the last child has to inform everyone what the original message was. Of course, the message is corrupted (one could argue that this is the goal of the game), since at every step in the cascade, there is a probability that the message is altered. However, beforehand, it is unclear what the final message precisely will be. Probably the final message will have a resemblance to the original input, but it will have changed. As importantly, repeating the same game with different children will lead to a different final message, but this final message will also be different from the final message of the first group of children. This variability in the final message, originating in the random changes of the message at each step of the cascade, is intrinsically present in every signaling cascade in the cell.

In living cells there are many of these signaling cascades. And, as important, many processes in cells are stochastic. The diffusion of molecules both extracellular and intracellular is a random process, but also the binding, unbinding, activation or deactivation dynamics of proteins are random. These cellular signaling cascades therefore suffer from the same problem as the telephone game; in every step there is a probability that the message is corrupted. The simple example of the children’s game is elegantly paraphrased by the data-processing
1.1 Cellular communication

inequality [5]. This inequality loosely states that

“the information between input and output in a sequence of steps can only go down, but never go up.”

Therefore, signal processing in cells is a complicated process, where, due to the stochasticity of the networks, it is difficult to maintain the original signal.

This problem becomes even more pronounced if we take into account the low-copy numbers which are typical for cellular signaling cascades. The central limit theorem states that fluctuations around the average response in any random process become insignificant with respect to the average response, if the average response becomes very large. However, the actual output of a signaling cascades can be very small, as small as only 10 copies of a specific molecule. Indeed, rhodopsin, the receptor for photons (light) in the eye can accurately measure the difference between 1 and 2 photons [6], eukaryotic cells can measure a concentration difference of order nanomolar (nM) between the front and back side [7], while gene transcription is sensitive to concentrations of nM, which in bacteria like E. coli corresponds to only a handful of molecules [8, 9].

The stochasticity leading to the inevitable uncertainty in the response of signaling cascades is a major challenge for cells. And yet, for the survival of a cell, an accurate response to incoming messages is important. An incorrect transfer of the measurement of a sugar concentration, could result in moving away from abundances of sugar, while the misinterpretation of the α-factor signal in yeast, prevents mating. To overcome these problems, signaling cascades have specific features.

Indeed, studies of complete protein (and gene) interaction networks have shown that specific architectures of coupled interactions are over-represented, and these are called network motifs [10, 11, 12]. The over-representation of these motifs has driven many studies, since this suggest that these motifs have an actual function, or at least an evolutionary advantage for cells [13, 14, 15, 16]. Let us consider what the evolutionary advantages of signaling cascades could be. Two advantages, or criteria, naturally arise: reliability and cost. Of course, many different advantages can be thought of, like transmission speed [17], memory-capability, both short-term [18] or long-term [19], robustness [20], change of temporal structure [21, 22] or timing [23, 24], but we argue that for a signaling cascade reliability and cost are essential.

Let us start with the second criterion: cost. As every process — e.g. gene expression, active transportation or phosphorylation — in the cells consumes energy, which is limited, it seems natural for a cell to try to reduce the energy consumption while transmitting signals through the cells. There are essentially two ways of reducing cost: 1) making less components and 2) share components between different tasks. Clearly, signaling cascades that rely on a small number of protein copies may have an evolutionary benefit, since they require a smaller production cost of making molecules. The other way of reducing the cost of maintaining a signaling cascade is to re-use the same proteins in multiple signaling cascades. Different signals then share a common cascade to transduce information. This principle, commonly referred to as multiplexing, is also at the heart of our modern macroscopic communication systems: indeed, many telephone calls are transmitted through the same wire.
and the European and American internet is connected through the main transatlantic glass fiber backbone. Both ways of reducing cost share an important drawback, the possible loss of correspondence between signal and response. In the case of the shared pathway there is the possibility of interference between the different signals that travel through this pathway, while in the case of the low copy number the correspondence can be lost due to the large impact of fluctuations. This directly brings us back to the first criterion: reliability.

Arguably, the most important property for a signaling cascade is to create a reliable connection between signal and response. Let’s think back of the telephone game one more time. To increase the likelihood of a perfect passing around of the message, instead of whispering, children could tell or, even more extreme, shout the message from one child to the next. This increase in amplitude of the voice in this game is similar to our definition of the gain of a signaling cascade. The gain is the amplification of the signal throughout the cascade and a large gain increases the reliability of the cascade as a whole. The counterpart of the gain is the noise: the uncertainty that is present in every biochemical process, be it protein production, activation, deactivation or degradation [25, 26, 27, 28, 29]. A large noise reduces the fidelity in the final message. In forthcoming sections, we will discuss the gain and the noise in much more detail.

The over-representation of these motifs suggests benefits for the cell, and since the gain and the noise are important characteristics of signaling cascades, indeed many studies have been devoted to the gain and noise properties of simple motifs [30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44]. However, as we stress here, to understand the reliability of signal transmission the gain and the noise should be studied simultaneously, since it is the relation, in fact their ratio, between the two that ultimately describes the information transmission capacity. Curiously, we have now introduced a new quantity, which might make sense to every reader intuitively, but, at first sight, seems hard to define quantitatively, namely information. Yet, it is precisely this quantity, information that quantifies the reliability of a signaling cascade. Therefore we now, in the next section, first discuss this in detail.

1.2 Information theory

In 1948 Claude Shannon published a paper which would shape the face of communication science for the rest of the century. In fact, one could state that that paper created a whole new scientific field: information theory [45]. An important part of information theory studies the reliability of communication systems. The challenge for every communication system is formulated by Shannon as

“\textit{The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.}”

Shannon was able to study this problem in a quantitative manner, which was an important achievement. To do this, he introduced and, importantly, quantified two important concepts, entropy and information.

Assume I have a very special dice, which has 8-sides, listing the numbers 1, 2, 3, \ldots, 8. I would challenge you for a guessing game, in which I roll this special dice, but I hide the outcome from you. How many closed (yes/no) questions do you need on average, to know the
outcome of the roll with absolute certainty? It is the answer to this question, that captures the quantification of information as defined by Shannon.

The correct answer is three questions. Assume I have rolled a 7. You start by asking if the outcome is larger than 4, thereby dividing the number of possible outcomes in two. The outcome is either 1, 2, 3, 4 or 5, 6, 7, 8. If my answer is yes, you apply the same trick by asking if the outcome is larger than 6, again dividing the set of possible outcomes in two equal halves. After your next question, you know, with absolute certainty that the outcome is 7. What is the relation between this process of "asking questions" and information?

In a mathematical framework, the outcome of the roll is a stochastic variable. Assuming I play fair and the dice is fair, all outcomes are as probable and the variable has a uniform probability distribution,

$$p(X = x) = \frac{1}{8}$$  \hspace{1cm} (1.1)

where $X$ is the stochastic variable and $x$ is one of the eight different outcomes. The entropy of the stochastic variable $X$, or more precisely, the entropy of the distribution $p(X)$
is defined as

$$H(X) = - \sum_{x=1}^{x=N} p(X = x) \log_2 p(X = x) \text{ bits}$$  \hspace{1cm} (1.2)$$

The entropy is often presented in units of bits, which is reflected by the use of \(\log_2\). The entropy can be defined for any probability distribution and is bounded by \(0 \geq H(X) \geq \log_2 N\). The entropy is zero if there is only a single possible outcome with probability \(p(X = x) = 1\). The entropy is maximal for a uniform distribution of the outcomes, such that all outcomes have equal probability, or equivalently all outcomes are as likely. The entropy is thus a measure for the uncertainty I have about the stochastic variable (Fig. 1.2).

For a continuous variable the entropy is ill-defined. From a mathematical perspective, the value of a continuous real variable has infinite precision. No matter how accurate your derivation is, or how many questions you ask, you will never discover the full precision of the number. For all practical purposes, continuous variables are discretized into bins and the entropy for the resulting discrete distribution is derived.

The entropy for our stochastic variable \(X\), the outcome of the dice roll,

$$H(X) = - \sum_{x=1}^{x=8} p(X = x) \log_2 p(X = x) = 3 \text{ bits}$$  \hspace{1cm} (1.3)$$

Strikingly, or maybe not, this exactly corresponds to the three questions you had to ask to obtain the answer. Phrased differently, and more precisely, this is the minimum number of questions required to obtain with absolute certainty the value of any stochastic number, on average. Of course, you could have guessed, and maybe even have guessed right, but on average guessing would have taken you more questions, than the bi-sectioning procedure described before [46].

1.2.1 Mutual information

The entropy is an important quantity in information theory. Instead of defining the entropy for \(X\) alone, we can also define the conditional entropy \(H(X|Y)\)

$$H(X|Y) = - \sum_{y} p(Y) \sum_{x} p(X|Y) \log_2 p(X|Y)$$  \hspace{1cm} (1.4)$$

with \(p(X) = p(X = x)\) and \(p(Y) = p(Y = y)\). The inner sum reflects the uncertainty in \(X\) for a specific value of \(Y\). If \(X\) and \(Y\) are independent, \(p(X, Y) = p(X)p(Y)\) and \(p(X|Y) = p(X)\), and as a result, the conditional entropy \(H(X|Y) = H(X)\). This indicates that \(Y\) is not informative on \(X\). However, if \(X\) and \(Y\) are not independent, \(H(X|Y) < H(X)\), and on average the uncertainty about \(X\) is reduced by knowledge of \(Y\). This insight provides the idea for the definition of a new quantity, the mutual information. The mutual information is a unique measure of interdependency between \(X\) and
1.2 Information theory

$Y$ [45]. Its definition is

$$I (X; Y) \equiv H (X) - H (X|Y)$$

$$= H (X) + H (Y) - H (X, Y)$$

$$= - \sum_{y} \sum_{x} p(X, Y) \log_2 p(X) + \sum_{y} \sum_{x} p(X, Y) \log_2 \frac{p(X, Y)}{p(Y)}$$

$$= \sum_{y} \sum_{x} p(X, Y) \log_2 \frac{p(X, Y)}{p(X)p(Y)}$$.

The mutual information is the difference of the entropy in $X$, $H (X)$, and the conditional entropy $H (X|Y)$, which is the uncertainty in $X$ given a specific $Y$ averaged over the probability distribution $p (Y)$. The mutual information reflects the decrease in the uncertainty of the variable $X$ upon knowledge of the value of the variable $Y$. More informally, the mutual information characterizes the information that is shared between $Y$ and $X$.

Takacik lists some important well-known properties of the mutual information [46]. First, the mutual information is symmetric in $X$ and $Y$, as is readily observed from Eq. 1.6 by interchanging $X$ and $Y$. The reduction in uncertainty in $X$ due to the knowledge of $Y$ is thus equivalent to the reduction in uncertainty of $Y$ upon knowledge of $X$. This is an intriguing property of the mutual information. Second, the mutual information can be defined for both discrete and continuous variables. Although the entropy for a continuous variable is ill-defined, the mutual information is not. Third, the mutual information obeys the data-processing inequality. For example for a cascade of three stochastic variables, where $X$ regulates $Y$ and $Y$ regulates $Z$, $X \rightarrow Y \rightarrow Z$, then $I (X; Z) \leq I (X; Y)$.

More in depth information on the mutual information can be found in the original paper by Shannon [45] or the books by Cover and Clover [5] or MacKay [47]. The mutual information as defined in Eq. 1.5 is used in Chapters 7 and 8.

In the neuroscience community information theory has been used for a few decades (see for a review [48]). But Ziv et al [49] were, arguably, the first to use information theory for a biochemical network, in this case a gene regulatory network. Since then many more studies have been performed, all using information transmission as the main quantity to describe the signaling properties of the network [50, 51, 52, 53, 54, 55, 56, 57, 58].

Noisy channel theorem

Let us take a look at the following process. We consider a system where Arthur J. (A) sends a single signal $S$ through a channel, where at the other end of the channel G. Bomans (B) receives a response $X$. A selects his signal from a probability distribution $p (S)$. If the channel is perfect, or in other words deterministic or noiseless, the response $X$ and signal $S$ are uniquely related and B would know precisely what $S$ is, if B has received $X$.

Most natural systems (unfortunately\footnote{Many studies have been devoted to the advantages of noise in natural systems, some of which are the introduction of heterogeneity in populations [59], stochastic amplification [60], stochastic resonance [61], increase in growth rate [62] or excitable systems [63]}) are not noiseless. As a result, the unique relation between $S$ and $X$ disappears, and B can only estimate what $S$ is using the response
For these channels, what is the maximum amount of information that can be transmitted? Thinking back of the telephone game, the information is transmitted more reliably if the participants would not whisper, but talk or even shout. In reality however, there is a physical limit to shouting, since it drains energy from anyone. Using this analogy loosely, any channel can transmit infinite information, but considering the cost of using energy, I constrain the average amplitude of the signal $A_S^2 \leq v$. With this knowledge alone, we can not determine the maximal information through the channel; for this we also require the noise characteristics of the channel. I assume that the noise $\eta$ in the channel has the three following properties: 1) $\eta$ has a Gaussian distributed amplitude ($\eta \sim \mathcal{N}(0, \sigma^2)$)

$$
\mathcal{N} \left( y; \mu, \sigma^2 \right) = p \left( y|\mu, \sigma^2 \right) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(y-\mu)^2}{2\sigma^2}},
$$

(1.9)

2) $\eta$ is spectrally white and 3) $\eta$ is independent of the signal $S$. This type of channel is called an additive Gaussian channel.

The maximal mutual information, defined as the channel capacity $C$, that can be transmitted through an additive Gaussian channel is

$$
C = \max I (S, X) = \frac{1}{2} \ln \left( 1 + \frac{A_S^2}{\sigma^2} \right) \text{ nats.}
$$

(1.10)

The channel capacity is a function of the ratio of the signal amplitude and the noise strength, which is commonly referred to as signal-to-noise-ratio (SNR). The channel capacity is obtained if the input distribution $p(S)$ is Gaussian distributed: $\mathcal{N} \left( 0, A_S \right)$ [47, 64]. This is an important observation, which I will rephrase: given a specific variance of the input distribution $p(S)$ and a Gaussian channel, the mutual information between $S$ and $X$ is maximized if $S$ has a Gaussian distribution. Very related to this observation is the following statement: for a channel with a Gaussian distribution for the input signal the mutual information is minimal if the conditional distribution between signal and response $p(x|s)$ is Gaussian [64].
The origin of this lower bound is that, for a given variance, a Gaussian distribution has the maximum entropy. Therefore, the channel with a Gaussian conditional distribution between signal and response “creates” the largest possible uncertainty between signal and response and therefore a lower bound on the mutual information.

**Mutual information for time-varying signals**

As we have seen the mutual information is a measure for the statistical dependence between two single univariate stochastic variables $X$ and $Y$. But a straightforward extension to this analysis provides the mutual information between two multivariate stochastic variables $V = (V_1, V_2, \ldots, V_N)^T$, $W = (W_1, W_2, \ldots, W_M)^T$ \cite{61, 64, 65, 66, 67}. This extension is the connection we require to obtain the mutual information between time-varying signals.

Assume we have a stochastic process which provides us with a continuous time trace $V(t)$. This trace has Gaussian fluctuations such that multiple realizations of the process are different, and we can write $V(t) = \langle V(t) \rangle + v(t)$, with $\langle V(t) \rangle$ the time-dependent average and $v(t)$ the fluctuations around the average.

We sample the signal $v(t) = V(t) - \langle V(t) \rangle$ at $N$ evenly spaced time points to obtain the vector $v = (v(t_0), v(t_1), v(t_2), \ldots, v(t_N))^T$ (we thus assume the fluctuations to be the signal). The correlation matrix $C$ for $v$ is defined as

$$C_{ij} = \langle v(t_i) v(t_j) \rangle = \int v(t_i) v(t_j) p(v) \, dv$$

Since the fluctuations in the signal are Gaussian, the joint probability distribution for the vector $v$ is

$$p(v) = \frac{1}{\sqrt{(2\pi)^N |C|}} e^{-\frac{1}{2}v^T C^{-1}v},$$

where $C^{-1}$ is the inverse of the correlation matrix and $|C|$ is the determinant of the correlation matrix. With the definition of the probability distribution for the vector $v$ (Eq. 1.12), we can define the differential entropy for the multivariate stochastic continuous variable $v$ following Eq. 1.2

$$H(v) = \int dv p(v) \ln p(v) \quad \text{[nats]},$$

Note that, for later convenience, we have changed the base of the logarithm, and as a result, the differential entropy has units [nats]. Further, we observe that the differential entropy $H(v)$ is not necessarily larger than zero. Since the joint probability distribution $p(v)$ for all the elements $v(t_i)$ has Gaussian statistics, we can actually obtain an analytical expression for the entropy $H(v)$. Since the correlation matrix $C$ is a real, symmetric matrix, we write

$$C^{-1} = \mathbf{Q}\mathbf{D}\mathbf{Q}^T,$$

with $\mathbf{Q}^{-1} = \mathbf{Q}^T$ and $\mathbf{Q}^{-1}\mathbf{Q} = \mathbf{I}$, \hspace{1cm} (1.14)

where $\mathbf{I}$ is the identity matrix. The matrix $\mathbf{Q}$ has the eigenvectors of $C^{-1}$ as column vectors.
The matrix $D$ is a diagonal matrix with the eigenvalues of $C$ along the diagonal. The matrix $Q$ also transforms the vector components $v_i$ into new variables $u_i : \Rightarrow u = Q^T v$. The differential entropy of the distribution $p(v)$ is

$$H(v) = -\int \frac{1}{\sqrt{(2\pi)^N |C|}} e^{-\frac{1}{2} u^T D^{-1} u} \ln \left( \frac{1}{\sqrt{(2\pi)^N |C|}} e^{-\frac{1}{2} u^T D^{-1} u} \right) du$$

(1.15)

$$= \frac{1}{\sqrt{(2\pi)^N |C|}} \int e^{-\frac{1}{2} u^T D^{-1} u} \left( \ln \left( \frac{1}{\sqrt{(2\pi)^N |C|}} \right) + \frac{1}{2} u^T D^{-1} u \right) du$$

(1.16)

Since $D$ is diagonal, we simplify Eq. 1.16 by substituting

$$\int e^{\frac{1}{2} u^T D u} du = \prod_{i=1}^{N} \int du_i e^{-\frac{1}{2} \lambda_i u_i^2} = \sqrt{(2\pi)^N} \prod_{i=1}^{N} \frac{1}{\lambda_i} = \sqrt{(2\pi)^N}$$

(1.17)

where $\lambda_i$ is the $i^{th}$ eigenvalue of $D$. Since $|D| = |C^{-1}| = 1/|C|$, we obtain

$$H(v) = \ln \left( \frac{1}{\sqrt{2\pi e} N \sqrt{|C|}} \right) = \ln \left( N \sqrt{\frac{2\pi e}{|D|}} \prod_{i=1}^{N} \sqrt{\lambda_i} \right)$$

(1.18)

Eq. 1.18 gives the differential entropy for a multivariate, Gaussian distribution. The derivation of Eq. 1.18 can easily be extended to the joint probability distribution of two multivariate, Gaussian processes $v$ and $w$. Introducing the new vector $z = (v, w)^T$ with correlation matrix $Z$

$$Z = \begin{pmatrix} C_{vv} & C_{vw} \\ C_{wv} & C_{ww} \end{pmatrix},$$

(1.19)

an equivalent derivation holds. As a result, since we have analytical expressions for the differential entropies $H(v), H(w)$ and $H(v, w)$, following Eq. 1.6 we write the mutual information between the multivariate, Gaussian distributions $v$ and $w$ as [67]

$$I_t (v; w) = H(v) + H(w) - H(v, w) = \frac{1}{2} \ln \left( \frac{|C_{vv}| |C_{ww}|}{|Z|} \right).$$

(1.20)

As a simple example let us focus on the situation where $N$, the number of time points, is 1. Then the vectors $v$ and $w$ reduce to $v(t_0), w(t_0)$ and the correlation matrices reduce to the instantaneous (co)variances, $\sigma_{ij} = \langle i(t_0) j(t_0) \rangle$, with $i = v, w$ and $j = v, w$. With
these following Eq. 1.20 we obtain the instantaneous mutual information

\[ I_t (v; w) = \frac{1}{2} \ln \left( \frac{\sigma_{vv}^2 \sigma_{ww}^2}{\sigma_{vw}^2 \sigma_{ww}^2 - \sigma_{vw}^4} \right) = \frac{1}{2} \ln \left( 1 - \frac{\sigma_{vw}^4}{\sigma_{vv}^2 \sigma_{ww}^2 - \sigma_{vw}^4} \right) \]  

(1.21)

\[ = \frac{1}{2} \ln \left( 1 + \frac{\sigma_{vw}^4}{\sigma_{vv}^2 \sigma_{ww}^2 - \sigma_{vw}^4} \right) \]  

(1.22)

Comparing Eq. 1.21 with Eq. 1.10 we observe that [67]

\[ \text{signal-to-noise} = \frac{\sigma_{vw}^4}{\sigma_{vv}^2 \sigma_{ww}^2 - \sigma_{vw}^4} = \frac{g^2}{N} \sigma_{vv}^2 \]  

(1.23)

where we have defined the instantaneous gain \( g^2 \equiv \sigma_{vw}^4/\sigma_{vv}^4 \) and noise \( N \equiv |Z|/\sigma_{vv} = \sigma_{ww}^2 - \sigma_{vw}^4/\sigma_{vv}^2 = \sigma_{ww}^2 - g^2 \sigma_{vv}^2 \). The noise therefore is that part of the variance in \( w \), that is not “explained” by the input variance (the signal) multiplied with the instantaneous gain \( g^2 \).

The mutual information rate

We are almost there. In this last section on information theory we derive the mutual information rate \( R(\omega) \), which is used extensively in Chapters 2 and 3.

In the previous section we have derived the general expression for the mutual information between multivariate, Gaussian distributions (Eq. 1.18) for time-discretized signals. However, biochemical signals are not discretized in time, but are continuous. Here we make the connection to continuous signals from the time-discretized signals. To simplify the results we assume stationarity, meaning that \( \langle V(t) \rangle = \langle V \rangle \) and \( \langle v(t) v(t') \rangle = \langle v(\tau) v(0) \rangle \), where \( \tau = t - t' \) and similar expressions for \( W \).

We sample the traces \( v(t) \) and \( w(t) \), with length \( T \), at \( N \) points with spacing \( \Delta \) and, again, assume that the multivariate joint distribution \( p(v, w) \) is Gaussian. We write the general eigenvalue equation for the correlation matrix (Eq. 1.11) [61]

\[ \sum_k C_{vv} (t_j, t_k) e^{-i \omega \eta kj \Delta} = \lambda_n e^{-i \omega \eta nj \Delta}, \]  

(1.24)

with \( \omega_0 = 2\pi/T \). Since we have assumed stationarity the elements of \( C_{vv} \) depend only on the absolute time difference \( |j - k| \), and therefore the correlation matrix has a Toeplitz structure. In the limit \( T \rightarrow \infty, N \rightarrow \infty \) this means that [5, 66]

\[ \lim_{N, T \rightarrow \infty} |C_{VV}| = \frac{1}{4 \pi T} \int_{\omega_{\text{max}}}^{\omega_{\text{max}}} \ln S_{vv} (\omega) d\omega, \]  

(1.25)

where \( S_{vv} (\omega) \) is the power spectrum of the signal \( v(t) \) and \( \omega_{\text{max}} = \pi/\Delta \) is the angular Nyquist frequency. With Eq. 1.25 we define the entropy rate \( h_G \) as

\[ h_G (V) = \lim_{N, T \rightarrow \infty} \frac{H_G}{N \Delta} \]  

(1.26)
Note that in the limit $\Delta \to 0$, the entropy rate $h_G$ is ill-defined. To obtain the mutual information rate, we require the joint entropy rate $h_G(V, W)$ (Eq. 1.6). Although (in general) the correlation matrix $Z$ is not Toeplitz, it is possible to obtain the mutual information rate in terms of the (cross-) power spectra \[61, 63, 67\]

$$\lim_{N \to \infty} \frac{I(v; w)}{N \Delta} = R(v; w) = -\frac{1}{4\pi} \int_{\omega_{\text{max}}}^{\omega_{\text{max}}} d\omega \ln \left[ 1 - \frac{|S_{vw}(\omega)|^2}{S_{vv}(\omega) S_{ww}(\omega)} \right]$$

(1.27)

$$\equiv \frac{1}{4\pi} \int_{\omega_{\text{max}}}^{\omega_{\text{max}}} d\omega \ln \left[ 1 + \frac{g^2(\omega)}{N(\omega)} S_{vv}(\omega) \right] \left[ \text{nats s}^{-1} \right].$$

(1.28)

The mutual information rate is the sum over independent Gaussian channels, where for each frequency the (co)variance is given by the (cross-)power spectrum. For independent Gaussian channels, the system should be linear, such that each input at a specific frequency only leads to an output at that same frequency. Since not every network is linear, a linearization procedure is required, and for this we use the Linear-Noise Approximation \[68\] (see the section: Linear-Noise Approximation). The mutual information rate is defined as a bit-rate, like the bit-rate of traditional modems. In Eq. 1.28 a gain and noise are defined again, but here these are frequency dependent.

$$g^2(\omega) \equiv \left| \frac{S_{vw}(\omega)}{S_{vv}(\omega)} \right|^2$$

(1.29)

$$N(\omega) \equiv S_{ww}(\omega) - g^2(\omega) S_{vv}$$

(1.30)

We note the similarity in structure between Eq. 1.21 and Eq. 1.27, showing that the mutual information rate in general is proportional to the signal-to-noise and/or gain-to-noise ratio. The important difference between these two ratios is discussed in Chapters 2 and 3.

An example

Let’s assume that we have a set of signals, which have a Gaussian distribution such that $p(S) = \mathcal{N}(\mu_S, \sigma^2_S)$. Each independent signal leads to an output distribution of responses, and the conditional distribution is Gaussian $p(X|S) = \mathcal{N}(\mu_{X|S}, \sigma^2_{X|S})$. The mutual information can now be calculated through Eq. 1.5, after specifying the (conditional) mean and variance. For reasons that will become clear shortly, we assume the following forms

$$\mu_S = \langle S \rangle = k \frac{\lambda}{\lambda}, \quad \sigma^2_S = k \frac{\lambda}{\lambda},$$

$$\mu_{X|S} = \langle X|S \rangle = \frac{\rho}{\mu(S)} \sigma^2_{X|S} = \frac{\rho}{\mu(S)}$$

(1.31)

Note here that although the distributions are Gaussian, the channel noise $\sigma^2_{X|S}$ depends on the signal $S$. This is thus not an additive Gaussian channel, since the noise depends on the signal, which can be observed in Fig. 1.3a,b from the widening of the distribution.
$p(S/\langle S \rangle, X/\langle X \rangle)$ for increasing $S$. In Fig. 1.4a we show the mutual information as function of $\mu$ for constant $\langle S \rangle = 100$, with $\rho = \lambda = 1$. For increasing $\mu$ the mutual information $I(S; X)$ decreases. The origin of this decrease is the decrease of the value of the conditional response $\langle X|S \rangle$, leading to an increase of the coefficient of variation (CV) $\sigma_{X|S}^2/\langle X|S \rangle^2$.

This is shown in Fig. 1.3a,b, where for two values of $\mu$ (resp. $\mu = 0.1, 1$), the probability distribution $p(S/\langle S \rangle, X/\langle X \rangle)$ (gray contour), the dose-response relation (dashed light gray), and most importantly the conditional distribution $p(X/\langle X|S \rangle |S = \langle S \rangle)$ (red, logscale) are shown. For larger $\mu$, the conditional distribution is much wider, reflecting the larger CV and therefore the smaller mutual information. We repeat that in Chapters 7 and 8 the mutual information for time-independent input signals is used.

**Figure 1.3:** The contourplot shows the joint probability distribution $p(S/\langle S \rangle, X/\langle X \rangle)$ for two different values of $\mu$. Both the conditional mean $\langle X|S \rangle$ and variance $\sigma_{X|S}^2$ scale with $\mu^{-1}$. The coefficient of variation (CV) $\sigma_{X|S}^2/\langle X|S \rangle^2$ therefore increases with $\mu$. Indeed, the conditional distribution $p(X/\langle X|S \rangle |S = \langle S \rangle)$ widens for an increase in $\mu$ (red line, logscale). Since the CV increases, the mutual information $I(S; X)$ decreases. a) $\mu=0.1$ s$^{-1}$ b) $\mu=1$ s$^{-1}$. Parameters: $k=100$ s$^{-1}$, $\lambda=\rho=1$ s$^{-1}$.

Having observed what the mutual information is for constant signals, we switch to the mutual information for time-varying signals. We take the, arguably, most simple signaling cascade, consisting of a signal $S$ and a response $X$, which is governed by the following biochemical reactions

$$\emptyset \xrightarrow{k_S} S, \quad S \xrightarrow{\lambda} \emptyset, \quad S \xrightarrow{\delta_S} S + X, \quad X \xrightarrow{\delta_X} \emptyset. \quad \text{(1.32)}$$

As input signal we take the variations $s(t) = S(t) - \langle S \rangle$ and the output is $x(t) = X(t) - \langle X \rangle$. This system is described by the following set of Langevin equations [69]

$$\frac{ds}{dt} = -\mu s(t) + \eta_s(t) \quad \text{(1.33)}$$

$$\frac{dx}{dt} = \rho s(t) - \mu x(t) + \eta_x(t). \quad \text{(1.34)}$$

Here $\eta_x(t)$ denotes the noise due to the production and degradation of $X$, while $\eta_s(t)$ is not a noise source in the usual sense, but rather defines the ensemble of input signals,
which is assumed to obey Gaussian statistics. They both have the following two properties
\( \langle \eta_{\alpha}(t) \rangle = 0 \), \( \langle \eta_{\alpha}(t)\eta_{\alpha}(t') \rangle = A_{\alpha}\delta(t - t') \), where \( \alpha = s, x \) and \( A_{\alpha} = 2\mu \langle S \rangle, A_{x} = 2\lambda \langle X \rangle \) [70] (see the section: Linear-Noise Approximation). For this simple system we calculate the
instantaneous mutual information (Eq. 1.21) and the mutual information rate (Eq. 1.27).

The steady-state responses are (compare with Eq. 1.31)

\[
\langle S \rangle = \frac{k}{\lambda}, \quad \langle X \rangle = \frac{\rho}{\mu} \langle S \rangle = \frac{\rho k}{\lambda \mu},
\]

and the steady-state (co)variances are

\[
\frac{\sigma_{ss}^2}{\langle S \rangle} = 1, \quad \frac{\sigma_{xx}^2}{\langle X \rangle} = 1 + \frac{\rho}{\mu + \lambda}, \quad \sigma_{sx}^2 = \frac{\rho \langle S \rangle}{\mu + \lambda},
\]

The instantaneous mutual information is (Eq. 1.21)

\[
I_t(s; x) = \frac{1}{2} \ln \left[ 1 + \frac{\sigma_{sx}^4}{\sigma_{ss}^2 \sigma_{xx}^2 - \sigma_{sx}^4} \right] \]

\[
= \frac{1}{2} \ln \left[ \frac{(\lambda + \mu)(\lambda + \mu + \rho)}{(\lambda + \mu)^2 + \lambda \rho} \right].
\]

The instantaneous mutual information gives the information between \( X \) at time \( t \) and \( S \) at
time \( t \). For the system as described in Eq. 1.32 the response time of \( X \) is set by \( \mu \). We first
discuss some interesting limits. In the limit \( \lambda \to \infty \) \( I_t(s, x) = 0 \), since for \( \lambda \to \infty \), every
variation in \( s \) will decay, before a change in \( x \) can be established. Indeed, the number of \( X \)
molecules produced in the lifetime of a single \( S \) molecule is \( \rho / \lambda \), and in the limit \( \lambda \to \infty \),
\( \rho / \lambda \to 0 \). In the opposite limit, \( \lambda \to 0 \), the variations in \( s \) decay very slowly, and the
instantaneous mutual information saturates \( I_t \propto (\mu + \rho) / \mu \), where the saturation value
depends on the relative production and degradation rates of \( X \). In the limit \( \rho \to \infty \), the
instantaneous mutual information reaches a saturation value \( I_t \propto (\mu + \lambda) / \lambda \). In this
limit, the production of \( X \) molecules increases and therefore the influence of fluctuations in
the copy number of \( X \) decreases. Further, in this limit, we observe that an increase in
\( \mu \), reflecting faster tracking of the signal, increases the saturation value. The instantaneous
mutual information is independent of \( k \), the production rate of the signal (as long as \( k > 0 \).
A change in \( k \) changes the mean level of \( \langle S \rangle \), but the timescale for the decay of variations in
\( s \) only depends on \( \lambda \), since the capacity of \( x \) to reliably track the variations in \( s \) depends only
on the dynamics. As a function of \( \mu \), the instantaneous mutual information has a maximum,
which is obtained for \( \mu_{\text{max}} = \sqrt{\lambda(\lambda + \rho)} \),

\[
I_{t,\text{max}}(s; x) = \frac{1}{2} \ln \left[ \frac{\lambda + \sqrt{\lambda(\lambda + \rho)}}{2\lambda} \right].
\]

For \( \mu \gg \lambda \), the intrinsic fluctuations in \( X \) are much more rapid than variations in \( X \) due to
variations in \( S \). As a result, the instantaneous statistics of \( X \) are those of a Poisson birth-
death process with constant input S and conditional variance $\sigma^2_{x|s} \approx \langle X \rangle \propto \mu^{-1}$. This increase in the noise is not compensated by a similar increase in the instantaneous gain $\sigma^4_{x\lambda}/\sigma^4_{s\lambda} \propto \mu^{-2}$. As a result the instantaneous mutual information scales with $\mu^{-1}$. In the opposite limit, $\mu \ll \lambda$, the response X is much slower than the variations in S, and the response integrates over the fluctuations in S. The actual value of $x(t)$ therefore does not reflect the instantaneous value of $s(t)$, but the time-average of $s(t)$ over a time-window $t_{\text{int}} = \mu^{-1}$; clearly, if $\mu \to 0$ all the variations in $s$ are integrated and $x$ will not respond to changes in $s$. In Fig. 1.4b the instantaneous mutual information is shown as function of $\mu$ for three different values of $\rho$, where $\lambda = 1 \text{ s}^{-1}$. First, the non-monotonic form with the maximum at $\rho_{\text{max}}$ is observed. Next, if $\rho \gg 1$, the mutual information is large, since this implies a larger amplification of $x$ following changes in the signal $s$.

We note here that the instantaneous mutual information and the mutual information for constant signals are not equivalent. The instantaneous mutual information is influenced by temporal correlations of the signal and the response, while the mutual information for constant signals is not. In the limit $\lambda \to 0$ and for equal conditional variances, the instantaneous mutual information approaches the constant mutual information (Fig. 1.4a, gray dashed line).

![Figure 1.4](image_url)

**Figure 1.4:** a) The difference between $I(S;X)$ (red solid) and $H(S)$ (red dashed), respectively the mutual information and the entropy of the input signal $S$, for a discrete joint Gaussian distribution $p(S,X)$. $I(S;X)$ decreases with increasing $\mu$, since $\text{CV} \sigma^2_{X|S}/\langle X \rangle^2 \propto \mu$. The maximal value of the mutual information $\max I(S;X) = H(S)$ is obtained in the limit $\mu \to 0$. The mutual information is equal to the instantaneous mutual information in the limit $\lambda \to 0$ (gray line). For small $\mu$ the instantaneous mutual information is larger, since the instantaneous mutual information is obtained assuming continuous signals and therefore the conditional differential entropy $H(s|x)$ can become negative, reflecting infinite precision on the value of $s$ given knowledge of $x$. b) The instantaneous mutual information $I_1(s;x)$ increases for $\rho \geq \lambda$ since this implies that the amplification of $x(t)$ by the variations in $s(t)$ is larger. In the limit $\lambda \to 0$ (red dashed, black solid) the input signal effectively is constant and $\sigma^2_{X|x} = \langle X \rangle$, $I_1(s;x) = 1/2 \ln(1+\rho/\mu) = 1/2 \ln(\sigma^2_{Xx}/\langle X \rangle^2)$. In the limit $\lambda \to \infty$ the variations in $s$ decay before these are transmitted to $x$ leading to a small instantaneous mutual information (gray dashed, dark red solid). For large $\mu$, the instantaneous mutual information scales as $\rho \lambda/\mu$. $\lambda, \rho$ in $\text{s}^{-1}$.

Finally, we look for this simple system at the mutual information rate $R(s,x)$ (Eq. 1.27).
the mutual information between complete trajectories $s(t), x(t)$. The mutual information rate $R$ is obtained through the calculation of the (cross-)power spectra, which are

\begin{align}
S_{SS} &= \frac{2k}{\omega^2 + \lambda^2}, \\
S_{XX} &= \frac{2\rho k}{(\omega^2 + \lambda^2) (\omega^2 + \mu^2)}, \\
S_{SX} &= \frac{2\rho k}{(\omega^2 + \lambda^2) (\omega + \mu)},
\end{align}

and we obtain for $R(s; x)$

\begin{equation}
\lim_{T \to \infty} R(s; x) = \frac{1}{4\pi} \int_{-\infty}^{\infty} d\omega \ln \left( \frac{\omega^2 + \lambda^2}{\omega^2 + \lambda^2 + \rho \lambda} \right) = \frac{\lambda}{2} \left( \sqrt{1 + \frac{\rho}{\lambda}} - 1 \right).
\end{equation}

Most striking about this result is that the mutual information rate does not depend on $\mu$, the degradation rate for the fluctuations in $X$. However reconsidering this observation, this follows from the fact that the degradation events in $X$ do not carry additional information on the actual state of $S$, given that the production events of $X$ are known [67], since the degradation only depends on the current state of $X$. We take a closer look at the frequency-dependent gain and noise for this simple system,

\begin{align}
g^2(\omega) &= \frac{|S_{SX}(\omega)|^2}{S_{SS}^2(\omega)} = \frac{\rho^2}{\omega^2 + \mu^2}, \\
N(\omega) &= S_{XX}(\omega) - \frac{|S_{SX}(\omega)|^2}{S_{SS}^2(\omega)} S_{SS}(\omega) = \frac{2\rho k}{(\omega^2 + \lambda^2) (\omega + \mu)},
\end{align}

which lead to a constant gain-to-noise ratio

\begin{equation}
\frac{g^2(\omega)}{N(\omega)} = \frac{\rho \lambda}{2k} = \frac{\rho}{\langle S \rangle}.
\end{equation}

The gain-to-noise ratio therefore does not depend on the dynamics of the signal $s(t)$, but only on the dynamics of the transmission of the signal to $X$. This is a property which we will explore in much more detail in Chapters 2 and 3.

Lastly, we look at the frequency-dependent gain from another perspective. Assume we have a signal $s(t)$ which is transferred through a black box, of which we have no knowledge. We do however assume that the black box has a noise-additive effect. Our main question is, whether we can infer from $x(t)$ the original signal $s(t)$. That is, after measuring the response $x(t)$, we estimate the signal $s_{est}(t)$. Following [48], we use a linear filter $h(t)$ to estimate $s_{est}(t)$ from $x(t)$. In the frequency domain a linear filter operation is given by

\begin{equation}
S_{est}(\omega) = H(\omega) X(\omega)
\end{equation}
Minimizing the least square error between the estimated signal $S_{\text{est}}$ and the true signal $S$ 
$$
\min \left[ \left( (S_{\text{est}} (\omega) - S (\omega))^2 \right) \right]
$$
leads to the following form for the filter \[48\]

$$
H (\omega) = \frac{S_{XX} (\omega)}{S_{XX} (\omega)},
$$
(1.48)

The noise is defined as the difference between the true signal and the estimated signal

$$
N_{\text{e}} (\omega) = S (\omega) - S_{\text{est}} (\omega)
$$

and with the noise the signal-to-noise ratio is defined as

$$
\frac{S_{\text{est}} S_{\text{est}} (\omega)}{N (\omega)} = \frac{\langle H (\omega) X (\omega) H^\dagger (\omega) X^\dagger (\omega) \rangle}{\langle N_{\text{e}} (\omega) N_{\text{e}}^\dagger (\omega) \rangle}
$$

(1.49)

\begin{align*}
    & = \frac{H^2 (\omega) S_{XX} (\omega)}{S_{SS} (\omega) + H^2 (\omega) S_{XX} (\omega) - \langle H \dagger (\omega) X \dagger (\omega) S (\omega) \rangle - \langle H (\omega) X (\omega) S^\dagger (\omega) \rangle} \\
    & = \frac{|S_{XS} (\omega)|^2}{S_{XX} (\omega) S_{SS} (\omega) - |S_{XS} (\omega)|^2},
\end{align*}
(1.50)

where $\dagger$ denotes the complex conjugate. The signal-to-noise ratio in Eq. 1.49, with the gain and noise as defined in respectively Eq. 1.44, Eq. 1.45, is equal to $g^2 (\omega)/N (\omega) S_{SS} (\omega)$. The signal-to-noise ratio is thus equivalent to the best estimate of the signal given the output at a specific frequency, or the reciprocal of the uncertainty in $S (\omega)$ given a particular $X (\omega)$.

1.3 Linear-Noise Approximation

As discussed in the previous sections, all biochemical processes are stochastic. A rigorous way to describe stochastic processes is by formulating a master equation for the underlying process \[68\], which describes the time-evolution of the full probability distribution of a process. Assume we have a simple birth-death process for a protein $X$, which is described by the following two processes

$$
X \xrightarrow{k_f} X + 1, \quad X \xrightarrow{k_b} X - 1
$$
(1.52)

where $X$ is the actual copy number of $X$. The master equation that describes the probability to observe $X$ proteins is

$$
\frac{dp (X)}{dt} = -(k_f + k_b X) p (X) + k_b (X + 1) p (X + 1) + k_f p (X - 1).
$$
(1.53)

The first term describes the two ways to leave the state $X$, either by the production or degradation of $X$, while the other two terms describe the two options to enter the state $X$. The master equation is a complete description to the underlying stochastic problem and therefore analytic solutions captures all the statistics of the process (see e.g. \[34\]). However,
analytical solutions are more often than not very difficult to obtain, mostly due to the fact that the processes are non-linear. Even the most simple enzymatic reaction

\[ S + E \xrightarrow{k_1 \over k-1} ES, \]
\[ ES \xrightarrow{k_2} E + P, \]

where S is the substrate molecule, E is the enzyme and P is the product, is described by a non-linear master equation. It is well-known that many biological processes have even stronger non-linear dependencies, like the dependence of the flagellar motor in E. coli on switching proteins [71] or the dependence of the expression level of the pGAL1-D12 promoter on its repressor TetR [72]. While non-linear processes show very rich and fascinating dynamics [73], they are difficult to solve analytically. In general many people resort to Stochastic Simulations Algorithms (SSA) as has been pioneered for chemical kinetics by Gillespie [74, 75]. A different approach is to look for approximations of the master equation that lead to analytically tractable solutions.

One of these solutions is the Linear-Noise Approximation (LNA), which we now discuss in more detail. The LNA provides us with a simple form for analytical approximations of the noise in any non-linear stochastic process. Here we will follow the derivation as introduced by Gillespie [69, 76], and directly apply it to an example.

We again start from the biochemical equations, as introduced in Eq. 1.32. We define the vector \( \mathbf{Y}(t) = (S(t), X(t)) \), where \( S(t), X(t) \) are the respective copy number of species \( S, X \) at time \( t \). Next we define the propensity functions \( a_j \), where \( a_j(y) \, dt \) is the probability that in time \([t, t + dt]\), given \( \mathbf{Y}(t) = y \) a reaction \( j \) of occurs, such that we have for our example

\[ a_0 = k, \quad a_1 = \lambda S(t), \quad a_2 = \rho S(t), \quad a_3 = \mu X(t) \]

Next to the propensity function we define the state-change stoichiometry vector \( \nu_j \), with \( \nu_{ji} \) is the change in \( Y_i \) given that reaction \( j \) occurs, such that we have \( \nu_0 = (1, 0)^T \), \( \nu_1 = (-1, 0)^T \), \( \nu_2 = (0, 1)^T \) and \( \nu_3 = (0, -1)^T \). With the propensity function and stoichiometry vector the chemical master equation can be formulated,

\[ Y_i(t + \tau) = y_{i,0} + \sum_{j=1}^{M} K_j(\mathbf{Y}(t), \tau) \nu_{ji}, \]

where \( K_j(\mathbf{Y}(t), \tau) \) is a random variable which reflects the number of events \( N \) in each reaction channel \( j \) in a time \( \tau \) and \( y_{i,0} \) is the value of \( Y_i \) at time \( t \). Note that Eq. 1.57 is exact, since \( \mathbf{Y}(t) \) is time-dependent. Now we make two independent assumptions to obtain a simpler expression for the random variables \( K_j(\mathbf{Y}(t), \tau) \). Suppose that we can define a time \( \tau > 0 \), which

1. is such that the propensity functions \( a_j(\mathbf{Y}(t)) \) within the time window \([t; t + \tau]\)
are approximately constant, meaning that the propensity functions can be written as \( a_j(y_t) \) where \( y_t \) is the system state at time \( t \), the start of the time window, and

2. is so long that a large number of events \( Q \gg 1 \) per reaction channel \( j \) occur in the time interval \([t, t + \tau]\)

We note that these two assumptions contradict each other, since the first assumption requires a small \( \tau \), while the second assumption requires a large \( \tau \). However, we argue, following [69], that for systems with large copy numbers a time \( \tau \) can be constructed that suffices both assumptions. Following the first assumption, the random variable \( K_j(y_t, \tau) \) follow a Poisson distribution since all reactions are independent, \( K_j(y_t, \tau) \propto P(a_j(y_t) \tau) \), where \( a_j(y_t) \tau \) is the average number of events in reaction channel \( j \). The second assumption allows us to write the Poissonian random variables as normal random variables, due to the law of large numbers. The mean and variance of these normal variables are equal, since they originate from Poisson random variables. These two assumptions allow us to write

\[
Y_i(t + \tau) = y_{i,t} + \sum_{j=1}^{M} \nu_{ji} N_j(a_j(y_t) \tau, a_j(y_t) \tau) \\
= y_{i,t} + \sum_{j=1}^{M} \nu_{ji} a_j(y_t) \tau \sum_{j=1}^{M} \nu_{ji} \sqrt{a_j(y_t) \tau} N_j(0, 1),
\]

since \( N_j(\mu, \sigma^2) = \mu + \sigma N_j(0, 1) \). In a final transition we rewrite \( \tau \to dt \), and introduce the noise source \( \Gamma_j(t) \), where \( \langle \Gamma_j(t) \rangle = 0 \) and \( \langle \Gamma_j(t) \Gamma_{j'}(t') \rangle = \delta(t - t')(j - j') \), to obtain the Chemical Langevin Equation (CLE)

\[
\frac{dY_i(t)}{dt} = \sum_{j=1}^{M} \nu_{ji} a_j(Y(t)) + \sum_{j=1}^{M} \nu_{ji} \sqrt{a_j(Y(t))} \Gamma_j(t).
\]

This leads for our example to

\[
\frac{dS(t)}{dt} = k - \lambda S(t) + \sqrt{k} \Gamma_0(t) - \sqrt{\lambda S(t)} \Gamma_1(t),
\]

\[
\frac{dX(t)}{dt} = \rho S(t) - \mu X(t) + \sqrt{\rho S(t)} \Gamma_2(t) - \sqrt{\mu X(t)} \Gamma_3(t).
\]

The CLE provides us with a clear description of the time-dependent noise strength (term(s) involving \( \Gamma_j \)), which not necessarily is linear. The CLE (Eq. 1.60) implies a corresponding
(forward) Fokker-Planck equation for the multivariate probability distribution \( p(S, X) \) \[69\]

\[
\frac{\partial p(y,t|y_0,t_0)}{\partial t} = -\sum_{i=1}^{N} \frac{\partial}{\partial y_i} \left[ \left( \sum_{j=1}^{M} \nu_{ji} a_j(y) \right) p(y,t|y_0,t_0) \right] + \frac{1}{2} \sum_{i=1}^{N} \frac{\partial^2}{\partial y_i^2} \left[ \left( \sum_{j=1}^{M} \nu_{ji}^2 a_j(y) \right) p(y,t|y_0,t_0) \right] + \sum_{i,i'=1,i<i'}^{N} \frac{\partial^2}{\partial y_i y_{i'}} \left[ \left( \sum_{j=1}^{M} \nu_{ji} \nu_{j'i'} a_j(y) \right) p(y,t|y_0,t_0) \right] , \quad (1.63)
\]

To continue, we make, following \([70, 77]\), two additional assumptions. The first we describe here, while the second is given below Eq. 1.64. The first assumption is that the ensemble average is independent of time: \( \langle S(t) \rangle = \langle S \rangle \). We derive the time-dependence of the stationary (co)variances \([70, 78, 79]\),

\[
\frac{dC}{dt} = AC + C^T A^T + B , \quad (1.64)
\]

where \( C \) is the covariance matrix with entries \( C_{ij} = \sigma_{ij}^2 \), \( A \) is the Jacobian matrix for the dynamics of the average process, with entries \( A_{ij} = \partial / \partial y_i \left[ \langle \Sigma_j \nu_{ji} a_j(\langle y \rangle) \rangle \right] \), and \( B \) a matrix which describes the diffusion (or random) terms, that depend on the stochastic events in the process, with entries \( B_{ik} = \Sigma_j \nu_{ji} \nu_{jk} a_j(\langle y \rangle) \). Importantly, in the derivation of both \( A \) and \( B \) we assume that at steady-state fluctuations around the steady-state can be ignored, which implies

\[
\langle SX \rangle = \langle S \rangle \langle X \rangle . \quad (1.65)
\]

We stress however that for the calculation of the (co)variances, Eq. 1.65 is not used. This assumption is in general not valid for non-linear systems and application of the LNA therefore should be done with care. However, it is well-known that the Linear-Noise Approximation is accurate as long as the copy numbers \( \gg O(10) \) \([49] \) and Fig. 1.5. Since both \( A \) and \( B \) now directly follow from the description of the process, Eq. 1.64 directly allows for the calculation of the variance of every component around steady-state. For our example (Eq. 1.32) we have

\[
A = \begin{pmatrix} -\lambda & 0 \\ \rho & -\mu \end{pmatrix} , \quad B = \begin{pmatrix} 2\lambda \langle S \rangle & 0 \\ 0 & 2\mu \langle X \rangle \end{pmatrix} . \quad (1.66)
\]

This leads to the (co)variances as given in Eq. 1.36 which is as expected, since our example system is linear. However, for non-linear systems, the LNA is not necessarily correct, as is shown in Fig. 1.5.
1.4 Numerical optimization

In this last in-depth section I describe two methods of computational optimization, which share many features, but are also somewhat different. Both methods are computational tools and are intensively used in these thesis, which merit a larger background description. But first a small general description as to why we use numerical optimization techniques. In many of the problems that are studied in this thesis the capacity to perform a specific function in a small minimal systems is assessed, like acting as a logic gate (Chapter 5) or multiplexing biochemical signals (Chapters 7 and 8). Although the studied systems are small, they comprise already a large parameter space (more than 10 parameters), which makes it difficult to, from scratch, analytically derive results. However, numerical results can be obtained by performing computational optimization techniques, where numerically the full parameter-space is explored and solutions to the problem at hand are returned. Ideally, one would like to perform brute-force calculations, scanning over all parameter space, but this, given the size of the parameter-space, is in generally unfeasible. Therefore, more elegant algorithms have been developed.

In general, in a multi-dimensional parameter space, the performance function (e.g. the capacity to multiplex) is a highly non-trivial function of all the parameters, showing multiple maxima (peaks) and minima (valleys). In both algorithms a system is a specific point in this parameter space, and each system is characterized by its performance function. The main problem of the optimization algorithms used in this thesis, simulated annealing and numerical evolution, is that they do not necessarily obtain the global minimum. Indeed, both algorithms can optimize the system for a local minimum, but not necessarily the global

![Figure 1.5: The Linear-Noise Approximation (LNA) (black circles) provides accurate predictions for the variance compared with full stochastic simulations (red circles). a) In a system, equal to Eq. 1.32, but with non-linear (cooperative) production of X S² X S² + X0, where S² is a dimer of two S molecules. The LNA (black circles) captures the noise levels in X very well at all copy numbers compared to the full stochastic system (red circles). Parameters: k=10 s⁻¹, λ=1 s⁻¹, ρ=20 s⁻¹, μ is varied to change the average response. b) In a birth-death system with non-linear (cooperative) degradation X S² X + X S² X0, the LNA (black circles) slightly underestimates the noise levels in X compared to the full stochastic system (red circles). Lines show analytical solutions (notes by Andrew Mugler). The use of the LNA should therefore be done with care. Parameters: k=400 s⁻¹, μ is varied to change the average response.](image)
minimum. Luckily, in many studies in this thesis, I am not necessarily interested in the global optimum, but in a sufficient optimum. In other words, for me it is satisfactory if a specific system can perform a specific function, but not if it is the best system for the specific function.

1.4.1 Simulated Annealing

The first algorithm is simulated annealing (SA). In this method, each point $p$ of the parameter space is analogous to a state of some physical system, and the function $E(p)$ (the performance function) to be minimized is analogous to the internal energy of the system in that state. The goal is to bring the system, from an random initial state in parameter space, to the state with the minimum possible energy (or lowest performance function). At each step, the SA algorithm compares its current state $p$ with a neighboring point $p'$, by comparing the function $E(p)$ and $E(p')$. Based upon the comparison of these two states, two options are possible:

1. If $E(p') < E(p)$ the new state $p'$ is accepted as a more optimal state
2. If $E(p') > E(p)$ the new state $p'$ is accepted following a probabilistic measure, where the likelihood to accept is depends on $\exp(-\beta(E(p') - E(p)))$, where $\beta$ is an inverse "temperature".

The idea is that, due to the probabilistic acceptance of less optimal states, the simulated annealing algorithm is able to overcome local minima and find the global minimum. The inverse temperature $\beta$, which is increased in time, allows for variation of the acceptance of large moves in the uphill direction, thereby stimulating or preventing the algorithm to leave the (local) minimum. The SA algorithm is used in Chapter 7.

1.4.2 Numerical Evolution

The second algorithm is Numerical Evolution. This algorithm in a sense follows the laws of natural evolution for optimization. Instead of the SA algorithm, which modeled a single system, this algorithm models the evolution of a population. The algorithm is based upon ideas of Wright and Fisher [80, 81]. Evolution occurs in discrete, synchronous steps, where the population size remains constant. At each step, each member of the population produces offspring in proportion to its fitness. Then, mutations occur, and the mutated offspring yield the population for the next step.

In general, we have a "population" of $S$ initial systems $s$ (parameter points in the parameter space). Each point has fitness $E(s)$, and the total fitness for the population is $F = \Sigma_{s} E(s)$. At each step, $S$ new systems ("offspring") are drawn from the distribution $p_{s} = E(s)/F$, which weights each system by its fitness. Each new system is then "mutated" by changing a single parameter $p$ in the parameter space, at which point selection starts again. Since initially the parameter space is randomly sampled and the assignment of offspring is probabilistic, one expects that the NE algorithm is less prone to converge into a local minimum; however, this is not guaranteed. More background information on this algorithm can be found in [82]. The NE algorithm is used in Chapters 5 and 8.
1.5 Scope of this thesis

The work in this thesis focuses on signal transmission in biochemical networks, with a strong emphasis on dynamical signals. In Chapters 2 and 3 we study the influence of network topology on the information transmission capacity of a specific network. Many recurring motifs are observed experimentally in signaling networks. We study the effect of each motif on the transmission of time-varying signals, the gain, noise and the information rate, which is proportional to the ratio of the signal-to-noise, all as a function of the frequency of the incoming signal. In Chapter 2 we focus on networks with feedback loops and autoregulation, while in Chapter 3 we study networks with a feed-forward architecture. These chapters give insight in how each motif transmits information as a function of frequency. In Chapter 4 we switch gears. Instead of studying the information transmission, we study the precision by which a cell can estimate the concentration by monitoring the occupancy state of a receptor. Indeed, this is an interesting problem; what is the fundamental lower limit on the precision of measuring chemical concentrations by time-averaging? The answer to this question is provided in a seminal paper written by Howard Berg and Edward Purcell [83], but since then however disputed in other publications. In this chapter, together with our co-workers from Japan, we reconsider this fundamental limit using a new analytical approach as well as computer simulations. In Chapter 5 we focus on actual information processing by single molecules. In recent years it has been shown that logic operations can be performed by single molecules [84]. We wonder whether a receptor, by variation of kinetic parameters, can perform all possible logic operations? While this question is interesting on evolutionary timescales, we then focus down onto signaling timescales. Is it possible for cells, by recombination of receptor monomers into dimers, to access all different logic operations. In Chapter 6 we return to one of the original observations, namely that signals are time-varying. This by itself is remarkable in the context of variability reduction, since intuitively one would expect that oscillating signals tend to increase the variability in a response, compared to constant signals. In Chapter 6 we study this question in more detail for a simple model of gene-regulation, pinning down the question whether indeed oscillating signals always increase the variability in a response as we naïvely expect.

The last two chapters are reserved for the topic reflected in the title of this thesis: Multiplexing Biochemical Signals. It is commonly observed that within cells different signals share a common signaling pathway. In other words, independent signals are transmitted through a common pathway, but ultimately they lead to their own unique response. This raises an interesting question: can cells simultaneously transmit signals through a common signaling network and yet, respond to each signal uniquely? We can think of this hypothesis with respect to the telephone system. In this telephone system multiple calls are transmitted through the same wire, but every independent caller-receiver combination can have a lively conversation without interference from any of the other calls in the network. In Chapter 7 we study the idea of multiplexing biochemical signals for signals that are constant in time. We discuss an encoding and decoding strategy that a cell could employ that allows for multiplexing, with absolute fidelity in the signal transmission for each signal. In Chapter 8 we extend the ideas of Chapter 7 by looking not only at signals that are constant in time, but also by studying signals that vary in time. Again we focus on possible encoding
and decoding strategies that allow multiplexing of biochemical signals. We study a possible multiplexing network in more detail, especially looking at the influence of interference from the different signals onto each other, and the influence of noise on the information transmission.