OPTIMAL CELLULAR PREDICTION

AGE TJALMA

This thesis was reviewed by:

prof.dr. F.J. Bruggeman	Vrije Universiteit Amsterdam
prof.dr. T.S. Shimizu	Vrije Universiteit Amsterdam
prof.dr. B.B. Machta	Yale University
dr. M.S. Bauer	Technische Universiteit Delft
dr. F.M. Berger	Universiteit Utrecht

The work described in this thesis was performed at AMOLF, Science Park 104, 1098 XG Amsterdam, The Netherlands, part of the Dutch Research Council (NWO). Funding for this work was received from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 885065). Financial support from the institute AMOLF for printing of this thesis is gratefully acknowledged.

Cover image: Age Tjalma, based on *Random Distribution of 40,000 Squares Using the Odd and Even Numbers of a Telephone Directory, 50% Blue, 50% Red* by François Morellet (1960).
Printed by: Ipskamp Printing, Enschede

ISBN 978-94-6473-770-7

A digital version of this thesis is available online at www.ub.vu.nl and ir.amolf.nl. The DOI of this thesis is https://doi.org/10.5463/thesis.1164. Printed copies can be obtained by request via library@amolf.nl.

VRIJE UNIVERSITEIT

OPTIMAL CELLULAR PREDICTION

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. J.J.G. Geurts,
volgens besluit van de decaan
van de Faculteit der Bètawetenschappen
in het openbaar te verdedigen
op donderdag 19 juni 2025 om 11.45 uur
in de universiteit

door

Age Jens Tjalma

geboren te Hoorn

promotor:	prof.dr. P.R. ten Wolde
promotiecomissie:	prof.dr. F.J. Bruggeman
	prof.dr. T.S. Shimizu
	prof.dr. B.B. Machta
	dr. M.S. Bauer
	dr. F.M. Berger

If life is getting better, enjoy it more;
if it is getting worse, don’t worry about it!

Howard Berg on *E. coli* chemotaxis.

CONTENTS

1

INTRODUCTION

Do you remember the last time you crossed a busy road? You may not have given it much thought, but before crossing, you likely observed oncoming cars, estimated their distance and speed, and used this information to predict whether it was safe to cross. The reason you could make this prediction is that the past, where the cars were and how fast they were moving, contained information about the future, where they would be when you reached the middle of the road. Your eyes enabled you to gather this information from the past and store it in your brain, though not without a cost. It took physical resources: proteins to maintain sensory cells and neurons, energy to send signals to the brain, and time to process the information. But what limits the accuracy of such predictions? Could you make better predictions with more neurons or a faster brain? These are the types of questions at the heart of this dissertation—although we won't be studying the human brain.

Remarkably, the ability to predict the future is not unique to humans or to animals; even single-celled organisms can anticipate environmental changes. These organisms live in highly dynamic environments, to which they must constantly adapt to survive. To this end, they use a range of response strategies, tailored to the nature of the changes they encounter. In highly regular conditions, such as the daily light cycle, cells use a biological clock from which they can tell the time and hence the state of the current and future environment [1, 2]. Conversely, in environments where changes are entirely unpredictable, cells can only resort to either a detect-and-respond strategy or a bet-hedging strategy, where different cells in a population are prepared for different environmental states [3]. The most fascinating response strategy, however, lies in between these two extremes: that of prediction. When environmental changes are somewhat regular, cells can start to anticipate upcoming changes, and initiate a response ahead of time.

Prediction can offer an evolutionary advantage by increasing cellular fitness [4]. Experiments have revealed that bacteria capable of predicting shifts in oxygen levels [5] or the arrival of nutrients [6] experience faster growth. Modeling further suggests that prediction can enhance the ability of bacteria to move towards nutrient-rich regions [7]. Yet, a predict-and-anticipate strategy is only useful if the cell can reliably predict the future on a timescale longer than the time needed to mount a response. This leads to two key questions: how accurately can cells predict the future, and what limits the accuracy of their predictions?

While the cell needs to predict the future environment, it can only sense the present and remember the past (Fig. 1.1A). Consequently, for a given amount of information the cell can store about the present and past signal, there is a maximum amount of information it can possibly have about the future [8, 9] (Fig. 1.1C-I). This *information bound* is determined by the temporal statistics, i.e. the inherent (un)predictability, of the environment [9, 10] .

How close cells can come to this bound depends on the design of the biochemical signaling network that they use to sense and process environmental signals (Fig. 1.1B). To maximize its capacity to predict the cell must use its memory effectively: it should extract only those characteristics from the present and past signal that are most informative about the future [7, 10–12]. Whether it can do so, is determined by the topology of the signaling network. Moreover, like any information processing device, biochemical networks require resources to be built and run. Molecular components are needed

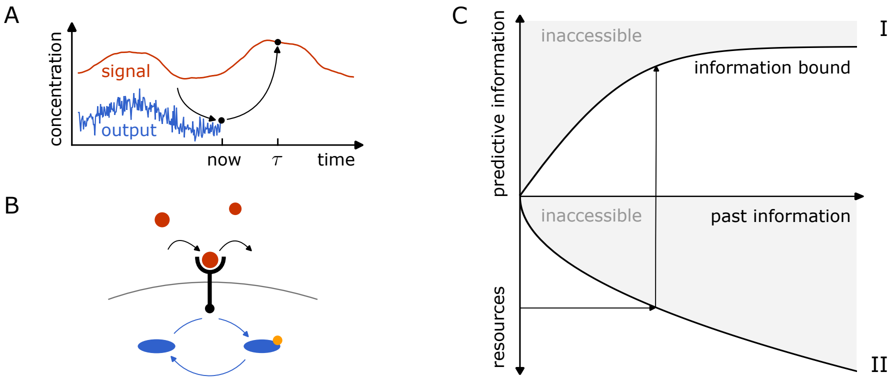


Figure 1.1: Cells use biochemical networks to remember the past and predict the future. (A) Cells compress the past input into the output of a signaling network from which the future input is then predicted. (B) The optimal topology of the network for predicting the future signal depends on the temporal statistics of the input signal. A push-pull network, as shown here, can optimally predict the future value of Markovian signals (Chapter 3). The push-pull network consists of a receptor that drives a downstream phosphorylation cycle, driven by ATP hydrolysis. As for any biochemical network, this comes at the cost of biophysical resources, such as proteins, energy, and time. (C) The predictive information on the future signal I_{pred} is fundamentally bounded by how much information I_{past} it has about the past signal (panel I), which in turn is limited by the resources necessary to build and operate the biochemical network (panel II) [8].

to construct the network, space is required to accommodate the components, time is needed to process the information, and energy is required to synthesize the components and operate the network [13]. Ultimately, these resources limit how much information the cell can collect from the past, and therefore they also limit its capacity to predict (Fig. 1.1C-II).

The fact that the costs and benefits of prediction are connected via the past and predictive information was first pointed out in the context of neuronal systems [8]. For these systems, tremendous progress has been made in characterizing the coding of information [12], and in quantifying the relation between past and predictive information [11]. However, understanding how the predictive information depends on resource cost is much harder for neuronal systems [14].

In contrast, cellular signaling systems provide a unique opportunity for revealing the resource requirements for prediction. For these systems we can readily quantify the information processing capacity as a function of the resources that are necessary to build and run them—protein copies, time, and energy [13, 15]. Moreover, cells use specialized signaling networks with distinct topologies, likely reflecting the typical temporal statistics of their input [7]. Cellular systems are thus ideal for elucidating the relationships between future and past information, system design (network topology), and resource constraints.

In Chapter 2 we use the information bottleneck method [9, 16] to derive the upper bound on the predictive information as a function of the past information for two classes of input signal. We find that, to reach the information bound for a Markovian input signal, the system should copy only the most recent signal value into its output. For a canonical class of non-Markovian signals our analysis shows that the optimal system must in general base its output on both the most recent signal value, and on its derivative.

In Chapters 3 and 4 we then investigate whether biochemical signaling networks exist that can reach the information bound for both classes of input signal. The ubiquitous push-pull motif (Fig. 1.1B) should be able to reach the bound for a Markovian input signal, because this network is at heart a copying device [13, 15, 17, 18]. Indeed, in Chapter 3 we find that this network can in principle reach the information bound. However, reaching the information bound is exceedingly costly because the energy required to drive the push-pull motif diverges as the copying speed increases. More surprisingly, even in a regime where this driving cost is negligible and the total resource cost is dominated by the cost of protein synthesis, the optimal system that maximizes the predictive information under a resource constraint is not at the information bound. The reason is that for a given resource availability, the system can increase both the past and the predictive information by moving away from the bound. This is a manifestation of a perhaps more general principle: the bits of past information that are most informative about the future, which are the most recent bits, are also the most costly.

Not all signals encountered by cells are Markovian. Living cells that navigate their environment typically experience signals that are influenced by their own persistent motion. Therefore, in Chapter 4, we study whether the *Escherichia coli* chemotaxis network can reach the information bound for a given regime of the non-Markovian signals studied in Chapter 2. We show that, to predict future concentration changes in a widely varying background concentration relative to the concentration change over the cell's orientational correlation time, the optimal system employs a perfectly adaptive, derivative-taking kernel. This is precisely the kernel that *E. coli* employs in its chemotaxis signaling network [19], and this system can therefore, in principle, reach the information bound. However, we again find that reaching the bound is prohibitively costly, because it requires taking an instantaneous derivative. The optimal system that maximizes the predictive information under a resource constraint is therefore not at the information bound, emerging from a trade-off between taking a derivative that is recent and one that is reliable. Finally, computing the past and predictive information directly from experimental data [20] reveals that the *E. coli* chemotaxis system indeed operates far from the information bound. Nevertheless, our analysis also indicates that the system is optimally tuned for the prediction of future concentration changes in shallow gradients under a resource constraint. This suggests that *E. coli* has evolved to use its resources for prediction most efficiently in shallow gradients.

The strategy of computing temporal derivatives by comparing the current signal to that in the recent past, over the so-called adaptation time, is pervasive in biology [19, 21–28]. But what limits the accuracy of such measurements and sets the optimal adaptation time remains unclear. In Chapter 5 we generalize the previously established sampling framework [13, 15, 29], and use it to study how the input statistics and resource availability determine the optimal design of the *E. coli* chemotaxis network. We find that an optimal adaptation time results from balancing the sampling error—a statistical error due to the stochasticity of the biochemical signaling network—and the dynamical error—a systematic error due to uninformative input variations. This balance shifts depending on resource availability and the steepness of the chemical gradients that the cell encounters. Both a larger resource availability and steeper gradients reduce the sampling error and allow for a shorter adaptation time, which in turn reduces the dynamical error.

Throughout all chapters we assume Gaussian statistics and use linear approximations [30]. This approach provides analytically tractable results and often allows a good approximation of the instantaneous mutual information between input and output of non-Gaussian systems [31–34]. However, the instantaneous mutual information may not fully capture the information that is encoded in the dynamics of the input and output trajectories. In Chapter 6, we apply a recent method [35] to compute the mutual information rate between trajectories exactly, and use the result to benchmark the previously established Gaussian approximation to the information rate [36, 37]. We find that the Gaussian approximation yields a lower bound to the true information rate of discrete linear systems. Moreover, for continuous nonlinear systems, the Gaussian approximation is accurate if the gain of the system is small. When the response of the system is slow relative to the input fluctuations the Gaussian approximation can also remain accurate for large gain, but only if the required two-point correlation functions are estimated directly from data.

The work in this dissertation improves our understanding of optimal strategies for prediction in general, and generates insight into how organisms may implement these strategies using biochemical signaling networks. Studying how close such networks are to optimality under different constraints reveals their most important limitations, and may hint at the evolutionary forces that have shaped them. Moreover, our work could be of use in the design of systems that need to process information under limited availability of physical resources, such as energy, components, or time.

2

ANY PREDICTION MUST BE BASED ON THE PAST

For any system to make a prediction, be it a human, a computer, or a bacterium, it needs to store information from the past into an internal output from which the future can be predicted. This means that the amount of information that is stored about the past, combined with the inherent uncertainty of the environment, fundamentally limits the predictive power. In this chapter, we derive this information bound for two types of input signal, and investigate the optimal mapping from past signal to internal output that allows the system to reach the bound. We find that for a simple Markovian input signal the optimal prediction device simply copies the most recent input. When including signal noise however, the optimal system should also incorporate values further in the past. For a non-Markovian signal which is defined by both its value and its derivative, the optimal mapping for the prediction of either the value or the derivative is one that combines both of these signal properties into a single output. When the objective is to predict both the signal value and its derivative, an optimal system maps the current input onto two distinct output components, given sufficient past information.

This chapter collects results from work published in Proceedings of the National Academy of Sciences **120**, e2303078120 (2023) [38] (optimal prediction of Markovian and non-Markovian signals); work done in collaboration with Jenny Poulton as first author (Section 2.4 Including signal noise [39]); and work done in collaboration with Vahe Galstyan as first author (Section 2.6 Predicting multiple components, in preparation).

Before we can answer the central question of this dissertation—whether cellular systems can predict optimally—we need to consider what optimal prediction entails. To make any prediction, a system must collect information about the future from the past. This so-called predictive information has to be encoded in an internal output from which the future can be predicted. Throughout this dissertation, we focus on the amount of predictive information that is encoded in the system's internal output. We leave the interesting, albeit difficult question of how to optimally decode such predictive information and translate it into behavior for future work.

Because all predictive information must be collected from the past, it is clear that the amount of information that a system collects from the past fundamentally limits the predictive information. Furthermore, the inherent uncertainty in the environment means that for an increasing amount of collected past information, the predictive information saturates. These two effects together yield an upper bound on the predictive information as a function of the past information. This so-called information bound can be determined using the Information Bottleneck Method [9]. Moreover, this method reveals the optimal mapping from the past input to the system output from which the future input is predicted. This is the mapping that extracts only those features from the past that are most predictive about the future, and therefore enables the system to reach the information bound.

In this chapter we first give a short introduction to the central measures from information theory used throughout this dissertation. Then we summarize the Information Bottleneck Method for Gaussian input signals [16], and subsequently apply it to a simple Markovian input signal generated by an Ornstein-Uhlenbeck process. We find that the optimal signaling system to predict such a signal must copy only the most recent input value into its output. When we however consider that the signal might be corrupted by high-frequency input noise, we find that the system should use inputs further in the past to average out the signal noise. In the sections that follow we turn our attention to a generic class of non-Markovian signals, generated by a stochastically driven damped harmonic oscillator. We first consider the prediction of a single property of this signal, either its value or its derivative. Our analysis reveals that, in general, the optimal system bases its output on both the current input value and derivative. How strongly each of these signal properties should contribute to the system output depends on the temporal statistics of the signal. Finally, we consider a scenario in which both the future concentration and derivative must be predicted. In this case, the optimal mapping is qualitatively different depending on the amount of past information that the system collects: for low past information the system should still combine the current input value and derivative into a single output component. However, as the past information increases a transition point is passed, after which the system should map the current input signal onto two distinct output components.

2.1. QUANTIFYING INFORMATION

Here we briefly discuss the core information theoretic quantities that we require in this work. For more detailed discussion and derivations see for example references [4, 40]. We quantify information using the so-called mutual information, established in the seminal work of Shannon [41]. The mutual information between two random variables measures to what extent knowing one of the variables reduces the uncertainty about the other. The uncertainty of any random variable can be quantified by its Shannon entropy,

$$H(X) = -\sum_x p(x) \log p(x), \tag{2.1}$$

The Shannon entropy is largest when a random variable is uniformly distributed, and decreases as its distribution becomes more sharply peaked, going to zero in the limit of a deterministic variable.

A highly intuitive definition of the mutual information between two random variables X and Y is one in terms of the Shannon entropy:

$$I(X;Y) = H(X) - H(X|Y), \tag{2.2}$$

which indeed shows that the mutual information is the reduction in uncertainty of X after learning Y. Using Eq. 2.1 and Bayes' theorem, $p(x,y) = p(x|y)p(y) = p(y|x)p(x)$, we can express the mutual information in terms of the marginal and joint probabilities of X and Y,

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{2.3}$$

This expression demonstrates the symmetry of mutual information: the reduction in uncertainty of X given Y is equivalent to the reduction in uncertainty of Y given X. Indeed, alternative expressions for the mutual information read

$$I(X;Y) = H(Y) - H(Y|X), \tag{2.4}$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \tag{2.5}$$

The entropy of Gaussian random variables is set by their variance. For an n-dimensional Gaussian variable $\boldsymbol{X}$ we have,

$$H(\boldsymbol{X}) = \frac{1}{2} \log\left((2\pi e)^n |\boldsymbol{\Sigma_x}|\right), \tag{2.6}$$

where $|\boldsymbol{\Sigma_x}|$ denotes the determinant of the covariance matrix $\boldsymbol{\Sigma_x}$. Using this Gaussian entropy in the various definitions of the mutual information above yields the following general expressions for the Gaussian mutual information

$$I(\boldsymbol{X};\boldsymbol{Y}) = \frac{1}{2} \log\left(|\boldsymbol{\Sigma_x}|/|\boldsymbol{\Sigma_{x|y}}|\right), \tag{2.7}$$

$$I(\boldsymbol{X};\boldsymbol{Y}) = \frac{1}{2} \log\left(|\boldsymbol{\Sigma_y}|/|\boldsymbol{\Sigma_{y|x}}|\right), \tag{2.8}$$

$$I(\boldsymbol{X};\boldsymbol{Y}) = \frac{1}{2} \log\left(|\boldsymbol{\Sigma_x}||\boldsymbol{\Sigma_y}|/|\boldsymbol{Z}|\right), \tag{2.9}$$

where $\boldsymbol{Z}$ is the covariance matrix of the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

2.2. INFORMATION BOTTLENECK METHOD

To compute the information bound and investigate the optimal mapping that maximizes the predictive information for a limited amount of past information we use the Information Bottleneck Method [9]. The objective function for the prediction of a future variable of interest $Z_\tau \equiv Z(t+\tau)$ is:

$$\max_{P(X_0|\boldsymbol{L}_p)} : \quad \mathcal{L} = I(X_0; Z_\tau) - \gamma I(X_0; \boldsymbol{L}_p). \tag{2.10}$$

The value of the sensing system output at the current time t_0 is $X_0 \equiv X(t_0)$. The variable of interest Z_τ at a future time $t_0+\tau$ is the signal property that the cell aims to predict, which can for example be a future signal concentration or its derivative, or even both. When the system of interest needs to predict a single signal characteristic, e.g. only the future concentration or only the future concentration derivative, one output component is sufficient for encoding the required information, as we describe in more detail below. Therefore, we first consider a scalar network output X. In section 2.6 we revisit this assumption. The vector $\boldsymbol{L}_p = (\delta\ell(0), \delta\ell(-\Delta t), \ldots, \delta\ell(-(N-1)\Delta t))^T$ is the past trajectory of ligand concentrations of length N, discretized with timestep Δt, where we have defined the vector in terms of deviations from the mean: $\delta\ell(t) \equiv \ell(t) - \bar{\ell}$. The mutual information between the current system output and the future property of interest is the predictive information $I_{\text{pred}} \equiv I(X_0; Z_\tau)$ [42, 43], and the mutual information between the current system output and the past ligand concentration trajectory is the past information $I_{\text{past}} \equiv I(X_0; \boldsymbol{L}_p)$. The Lagrange multiplier γ constrains the past information, i.e. it sets the compression level. Given a value of γ, Eq. 2.10 is maximized by optimizing the mapping of the past ligand concentration trajectory $\boldsymbol{L}_p$ onto the current output X_0. Since we have $I_{\text{past}} \geq I_{\text{pred}}$, for $\gamma = 1$ the objective function is maximized by $I_{\text{past}} = I_{\text{pred}} = 0$. As γ is decreased both the past and predictive information increase, and the parametric curve in the $I_{\text{past}} - I_{\text{pred}}$ plane that arises is the information bound (Fig. 1.1C-I). For $\gamma = 0$ there is no compression, and the past information is allowed to diverge. The predictive information is then only limited by the amount of information contained in the past about the future signal property: $I_{\text{pred}} \leq I(\boldsymbol{L}_p; Z_\tau)$.

GAUSSIAN INFORMATION BOTTLENECK

In general equation 2.10 can be difficult to solve, as all mappings from $\boldsymbol{L}_p$ to X_0 are allowed. However, the problem becomes analytically tractable when the joint probability distribution of $\boldsymbol{L}_p$ and Z_τ is a multivariate Gaussian. Here, we follow the procedure of Chechik and coworkers to obtain the optimal mapping from $\boldsymbol{L}_p$ to X_0 [16]. In the Gaussian model, this optimal mapping is a linear one [16]

$$X_0 = \boldsymbol{A}\boldsymbol{L}_p + \xi; \quad \xi \sim N(0, \sigma_\xi^2), \tag{2.11}$$

where $\boldsymbol{A}$ is a row vector which determines how strongly each entry in $\boldsymbol{L}_p$ contributes to the scalar output X_0 at any point in time. The random variable ξ is the noise added to the signal due to the stochastic nature of the mapping; it is a Gaussian random variable independent of $\boldsymbol{L}_p$ with 0 mean and variance σ_ξ^2. Finding the optimal mapping from $\boldsymbol{L}_p$ to X_0 corresponds to finding the optimal combination of $\boldsymbol{A}$ and σ_ξ^2. It can be shown that for any pair $(\boldsymbol{A},\sigma_\xi^2)$, there exists a pair $(\boldsymbol{A}',1)$ which yields the same values for I_{past} and I_{pred} after maximization of Eq. 2.10 [16].

To obtain the information bound, we rewrite Eq. 2.10 using the mutual information between Gaussian random variables (Eq. 2.7):

$$\mathcal{L} = \frac{1}{2}\log(\sigma_x^2/\sigma_{x|z_\tau}^2) - \gamma\frac{1}{2}\log(\sigma_x^2/\sigma_{x|L}^2), \tag{2.12}$$

with the total variance σ_x^2 in the output X_0, the output variance conditional on the future signal property $\sigma_{x|z_\tau}^2$, and the output variance conditional on the complete history of ligand concentrations $\sigma_{x|L}^2 \equiv \sigma_{x|\boldsymbol{L}_p}^2$. The latter is just the variance caused by the intrinsic noise, $\sigma_{x|L}^2 = \sigma_\xi^2 = 1$. The total variance in X_0 can be expressed in terms of the mapping vector $\boldsymbol{A}$ and the variance in the past signal using Eq. 2.11, $\sigma_x^2 = \boldsymbol{A}\boldsymbol{\Sigma}_L\boldsymbol{A}^T + 1$, where $\boldsymbol{\Sigma}_L \equiv \boldsymbol{\Sigma}_{\boldsymbol{L}_p}$ is the covariance matrix of the past ligand concentration trajectory $\boldsymbol{L}_p$. To express the output variance conditional on the future signal property Z_τ we use the Schur complement formula, which in general form reads:

$$\boldsymbol{\Sigma}_{x|y} = \boldsymbol{\Sigma}_x - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{yx}, \tag{2.13}$$

where $\boldsymbol{\Sigma}_{yx} = \boldsymbol{\Sigma}_{xy}^T$. Using this formula to rewrite $\sigma_{x|z_\tau}^2$, and then using the linear relation from Eq. 2.11 again, we obtain $\sigma_{x|z_\tau}^2 = \boldsymbol{A}\boldsymbol{\Sigma}_{L|z}\boldsymbol{A}^T + 1$.

Filling in the expressions for the variances in $\mathcal{L}$ (Eq. 2.12) yields,

$$\mathcal{L} = \frac{1}{2}\left((1-\gamma)\log\left(\left|\boldsymbol{A}\boldsymbol{\Sigma}_L\boldsymbol{A}^T + 1\right|\right) - \log\left(\left|\boldsymbol{A}\boldsymbol{\Sigma}_{L|z}\boldsymbol{A}^T + 1\right|\right)\right). \tag{2.14}$$

For any symmetric matrix $\boldsymbol{C}$ we have $\frac{\delta}{\delta A}\log|\boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^T| = \left(\boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^T\right)^{-1}2\boldsymbol{A}\boldsymbol{C}$, such that we obtain for the derivative of $\mathcal{L}$ to $\boldsymbol{A}$,

$$\frac{\delta\mathcal{L}}{\delta\boldsymbol{A}} = (1-\gamma)\frac{\boldsymbol{A}\boldsymbol{\Sigma}_L}{\boldsymbol{A}\boldsymbol{\Sigma}_L\boldsymbol{A}^T+1} - \frac{\boldsymbol{A}\boldsymbol{\Sigma}_{L|z}}{\boldsymbol{A}\boldsymbol{\Sigma}_{L|z}\boldsymbol{A}^T+1}. \tag{2.15}$$

In our case $\boldsymbol{A}$ is a row vector, and both denominators are thus scalars. We find the maximum of $\mathcal{L}$ by equating its derivative (Eq. 2.15) to 0, yielding

$$\boldsymbol{A}\boldsymbol{\Sigma}_{L|z}\boldsymbol{\Sigma}_L^{-1} = (1-\gamma)\frac{\boldsymbol{A}\boldsymbol{\Sigma}_{L|z}\boldsymbol{A}^T+1}{\boldsymbol{A}\boldsymbol{\Sigma}_L\boldsymbol{A}^T+1}\boldsymbol{A}. \tag{2.16}$$

For this equality to hold $\boldsymbol{A}$ must either be identically 0, or a left eigenvector of the matrix $\boldsymbol{\Sigma}_{L|z}\boldsymbol{\Sigma}_L^{-1}$ with eigenvalue:

$$\lambda = (1-\gamma)\frac{\boldsymbol{A}\boldsymbol{\Sigma}_{L|z}\boldsymbol{A}^T+1}{\boldsymbol{A}\boldsymbol{\Sigma}_L\boldsymbol{A}^T+1}. \tag{2.17}$$

Here, we note that if the signal statistics is sufficiently rich and the prediction complexity sufficiently large (because, for example, multiple signal characteristics need to be predicted), then the matrix $\boldsymbol{\Sigma}_{L|z}\boldsymbol{\Sigma}_L^{-1}$ has multiple eigenvectors with non-trivial eigenvalues $0 < \lambda_i < 1$ [16]. This reflects the idea that storing the past information that is necessary to enable this complex prediction task may require multiple output components, i.e. an output vector $\boldsymbol{X}$, where each output component has an integration kernel given by one of the eigenvectors of $\boldsymbol{\Sigma}_{L|z}\boldsymbol{\Sigma}_L^{-1}$ [16]. However, for Markovian signals only one eigenvector with non-trivial eigenvalue $0 < \lambda < 1$ emerges, which means that one output component is sufficient to encode the required information. For the non-Markovian signals studied below, $\boldsymbol{\Sigma}_{L|z}\boldsymbol{\Sigma}_L^{-1}$ has two eigenvectors if both the future value and its derivative need to be predicted; to optimally predict both features from the current output, two output components are then required, provided I_{past} is sufficiently large. We reveal an intuitive interpretation of the transition point after which an additional component is required for optimal prediction in section 2.6. However, throughout this dissertation we generally consider the scenario that only one signal feature needs to be predicted, in which case only one non-trivial eigenvector emerges, and one output component is sufficient for encoding the required information.

We can define the optimal mapping $\boldsymbol{A} = ||A||\boldsymbol{v}$ where $\boldsymbol{v}$ is the normalized left eigenvector of $\boldsymbol{\Sigma}_{L|z}\boldsymbol{\Sigma}_L^{-1}$ corresponding to its smallest eigenvalue, $0 < \lambda < 1$. The magnitude can be found by solving Eq. 2.17 for $||A||$, using from Eq. 2.16 that $\lambda\boldsymbol{v}\boldsymbol{\Sigma}_L\boldsymbol{v}^T = \boldsymbol{v}\boldsymbol{\Sigma}_{L|z}\boldsymbol{v}^T$. We obtain for the optimal mapping,

$$\boldsymbol{A}^{\text{opt}} = \begin{cases} \sqrt{\frac{1-\gamma-\lambda}{\boldsymbol{v}_1\boldsymbol{\Sigma}_L\boldsymbol{v}_1^T\lambda\gamma}}\boldsymbol{v}_1 & \text{for} \quad 0 < \lambda < 1-\gamma, \\ 0 & \text{for} \quad 1-\gamma \leq \lambda \leq 1. \end{cases} \tag{2.18}$$

We can substitute $||A||^2 = (1-\gamma-\lambda)/(\boldsymbol{v}\boldsymbol{\Sigma}_L\boldsymbol{v}^T\lambda\gamma)$ in the definitions for the mutual information (see for example Eq. 2.12 where the first RHS term is the predictive information and the second RHS term is the past information) to express them in terms of λ and γ. For the past information we obtain:

$$\begin{aligned} I_{\text{past}}(\boldsymbol{L}_p; X_0) &= \frac{1}{2}\log\left(||A||^2\boldsymbol{v}\boldsymbol{\Sigma}_L\boldsymbol{v}^T + 1\right), \\ &= \frac{1}{2}\log\left(\frac{1-\gamma}{\gamma}\frac{1-\lambda}{\lambda}\right). \end{aligned} \tag{2.19}$$

And for the predictive information:

$$\begin{aligned} I_{\text{pred}}(X_0; Z_\tau) &= \frac{1}{2}\log\left(||A||^2\boldsymbol{v}\boldsymbol{\Sigma}_L\boldsymbol{v}^T + 1\right) - \frac{1}{2}\log\left(||A||^2\boldsymbol{v}\boldsymbol{\Sigma}_{L|\ell_\tau}\boldsymbol{v}^T + 1\right), \\ &= I_{\text{past}} - \frac{1}{2}\log\left(\frac{1-\lambda}{\gamma}\right), \\ &= \frac{1}{2}\log\left(\frac{1-\gamma}{\lambda}\right). \end{aligned} \tag{2.20}$$

2.3. MARKOVIAN INPUT

In this section we consider a Markovian signal and derive the optimal mapping from the past ligand concentration trajectory $\boldsymbol{L}_p$ to the current output X_0. For the ligand concentration dynamics we use a 1-dimensional OU-process

$$\delta\dot{\ell} = -\delta\ell/\tau_\ell + \eta_\ell(t), \tag{2.21}$$

where the ligand concentration is defined in terms of the deviation from its mean $\delta\ell = \ell(t) - \bar{\ell}$. The correlation time is given by τ_ℓ, and the noise $\eta_\ell(t) \equiv \sigma_\ell\sqrt{2/\tau_\ell}\xi(t)$ is derived from a unit white noise process $\xi(t)$, such that $\langle\eta_\ell(t)\eta_\ell(t')\rangle = 2\sigma_\ell^2/\tau_\ell\delta(t-t')$. The concentration deviation at time t is thus given by

$$\delta\ell(t) = \int_{-\infty}^{t} dt' e^{-(t-t')/\tau_\ell}\eta_\ell(t'), \tag{2.22}$$

where the initial condition is forgotten because we consider the stationary regime. We obtain for the steady-state auto-correlation

$$\begin{aligned}
\langle\delta\ell(\tau)\delta\ell(0)\rangle &= \frac{2\sigma_\ell^2}{\tau_\ell}\int_{-\infty}^{\tau} dt' \int_{-\infty}^{0} dt e^{-(\tau-t')/\tau_\ell} e^{t/\tau_\ell}\delta(t-t'), && (2.23)\\
&= \frac{2\sigma_\ell^2 e^{-\tau/\tau_\ell}}{\tau_\ell}\int_{-\infty}^{0} dt e^{2t/\tau_\ell}, && (2.24)\\
&= \sigma_\ell^2 e^{-\tau/\tau_\ell}. && (2.25)
\end{aligned}$$

To obtain the information bound for prediction of the future ligand concentration of a Markovian signal, we need to determine the eigenvalues and vectors of the matrix (see Eqs. 2.16 and 2.17)

$$\boldsymbol{W} = \boldsymbol{\Sigma}_{L|\ell_\tau}\boldsymbol{\Sigma}_L^{-1}. \tag{2.26}$$

Using the Schur complement formula (Eq. 2.13) to rewrite the conditional matrix gives $\boldsymbol{\Sigma}_{L|\ell_\tau} = \boldsymbol{\Sigma}_L - \boldsymbol{\Sigma}_{L\ell_\tau}\boldsymbol{\Sigma}_{L\ell_\tau}^T/\sigma_\ell^2$. Then defining the normalized matrices $\boldsymbol{R}_L = \boldsymbol{\Sigma}_L/\sigma_\ell^2$ and $\boldsymbol{R}_{L\ell_\tau} = \boldsymbol{\Sigma}_{L\ell_\tau}/\sigma_\ell^2$ we find

$$\boldsymbol{W} = \mathbb{I}_N - \boldsymbol{R}_{L\ell_\tau}\boldsymbol{R}_{L\ell_\tau}^T\boldsymbol{R}_L^{-1}. \tag{2.27}$$

where N is the length of the input trajectory $\boldsymbol{L}_p$. The correlation matrix of the past trajectory is symmetric with entries $\boldsymbol{R}_L^{(i,j)} = \exp(-|i-j|\Delta t/\tau_\ell)$, where Δt is the discretization timestep of the past trajectory $\boldsymbol{L}_p$ and i ranges from 1 to N. This is a Kac-Murdock-Szegö matrix, and its inverse is known:

$$\boldsymbol{R}_L^{-1} = \frac{1}{1-e^{-2\Delta t/\tau_\ell}}\begin{pmatrix}
1 & -e^{-\Delta t/\tau_\ell} & 0 & \cdots & \cdots & 0\\
-e^{-\Delta t/\tau_\ell} & 1+e^{-2\Delta t/\tau_\ell} & -e^{-\Delta t/\tau_\ell} & \cdots & \cdots & 0\\
0 & -e^{-\Delta t/\tau_\ell} & 1+e^{-2\Delta t/\tau_\ell} & \ddots & \cdots & 0\\
\vdots & \vdots & \ddots & \ddots & \ddots & \vdots\\
0 & \cdots & \cdots & -e^{-\Delta t/\tau_\ell} & 1+e^{-2\Delta t/\tau_\ell} & -e^{-\Delta t/\tau_\ell}\\
0 & \cdots & \cdots & 0 & -e^{-\Delta t/\tau_\ell} & 1
\end{pmatrix}. \tag{2.28}$$

Note that this inverse matrix is tridiagonal. The length N cross-correlation vector between past trajectory and future concentration has entries $\boldsymbol{R}_{L\ell_\tau}^{(i)} = \exp(-(\tau+(i-1)\Delta t)/\tau_\ell)$.

The product of the correlation matrices in Eq. 2.27 turns out to be surprisingly simple:

$$\boldsymbol{R}_{L\ell_\tau}\boldsymbol{R}_{L\ell_\tau}^T\boldsymbol{R}_L^{-1} = e^{-2\tau/\tau_\ell}\begin{pmatrix} 1 & 0 & \dots & 0 \\ e^{-\Delta t/\tau_\ell} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ e^{-(N-1)\Delta t/\tau_\ell} & 0 & \dots & 0 \end{pmatrix}. \tag{2.29}$$

Using this result we can straightforwardly determine the eigenvalues,

$$\begin{aligned} &|\boldsymbol{W} - \lambda\mathbb{I}_N| = 0, \\ &\left|\begin{pmatrix} 1-\lambda-e^{-2\tau/\tau_\ell} & 0 & \dots & 0 \\ -e^{-(\tau+\Delta t)/\tau_\ell} & 1-\lambda & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -e^{-(\tau+(N-1)\Delta t)/\tau_\ell} & 0 & \dots & 1-\lambda \end{pmatrix}\right| = 0. \end{aligned} \tag{2.30}$$

The only contribution to the determinant comes from the diagonal, and the only non-trivial eigenvalue is thus $\lambda = 1 - e^{-2\tau/\tau_\ell}$, revealing that the optimal mapping is onto a one-dimensional scalar output X_0. The corresponding left eigenvector is given by

$$\boldsymbol{v}_1\boldsymbol{W} = (1 - e^{-2\tau/\tau_l})\boldsymbol{v}_1, \tag{2.31}$$

which holds for $\boldsymbol{v}_1 = \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix}$. The optimal mapping for the prediction of a one-dimensional OU-process is therefore to copy its most recent value. This agrees with intuition as for any Markovian process, all the information about the future signal is contained in the most recent value. For a continuous input signal (rather than a discretized signal), and a continuous integration kernel $k(t)$ (rather than a mapping vector $\boldsymbol{A}$), this means that the optimal integration kernel is $k^{\text{opt}}(t) = a\delta(t)$. In Chapter 3 we will see that there indeed exist biochemical networks that perform such a copy operation.

2.4. INCLUDING SIGNAL NOISE

In the previous section we have studied the optimal mapping for the prediction of a Markovian input signal. Because the signal is Markovian, any information about the future state of the signal is contained in its most recent value. In terms of the mutual information between the past and future signal, which sets the upper bound on the predictive information, we have $I(\boldsymbol{L}_p;\ell_\tau) = I(\ell_0;\ell_\tau)$. Therefore, the optimal mapping acts on the most recent signal value $\ell(t_0) = \ell_0$. However, this result requires that the system can map the true current input signal ℓ_0 independently onto an output which is corrupted by white noise. In reality, a signal measurement is likely to be noisy already at the input level of the signaling system, for example due to stochastic sampling of the true signal. In this scenario, the system only has access to a degraded input signal, and the mapping that the system performs acts on this degraded signal, comprising both the true signal and additional input noise.

To elucidate the effect of input noise on the optimal system design, we here consider a degraded input signal $\boldsymbol{S}_p$, which at time t is given by

$$s(t) = \ell(t) + \zeta(t), \tag{2.32}$$

where $\ell(t)$ represents the true input signal obeying Markovian dynamics as in Eq. 2.21, with a correlation function given by Eq. 2.25. The input noise $\zeta(t)$ has zero mean and exponentially decaying correlations over time $\langle\zeta(t)\zeta(t')\rangle = \sigma_\ell^2\sigma_\zeta^2\exp(-t/\tau_\zeta)$, with decay rate τ_ζ^{-1} and relative noise strength σ_ζ.

The objective remains the same, to maximize the predictive information $I(X_0;\ell_\tau)$ under constrained past information $I(X_0;\boldsymbol{L}_p)$ (Eq. 2.10). However, the system kernel maps the degraded past input signal to the output instead of the underlying true signal,

$$X_0 = \boldsymbol{A}\boldsymbol{S}_p + \xi = \boldsymbol{A}\big(\boldsymbol{L}_p + \boldsymbol{\zeta}_p\big) + \xi. \tag{2.33}$$

The different variances that enter the Lagrangian (Eq. 2.12) are now functions of the mapping $\boldsymbol{A}$, the variance of the true signal, the input noise, and the intrinsic noise,

$$\sigma_x^2 = \boldsymbol{A}\big(\boldsymbol{\Sigma}_L + \boldsymbol{\Sigma}_\zeta\big)\boldsymbol{A}^T + \sigma_\xi^2, \tag{2.34}$$

$$\sigma_{x|L}^2 = \boldsymbol{A}\boldsymbol{\Sigma}_\zeta\boldsymbol{A}^T + \sigma_\xi^2, \tag{2.35}$$

$$\sigma_{x|\ell_\tau}^2 = \boldsymbol{A}\big(\boldsymbol{\Sigma}_{L|\ell_\tau} + \boldsymbol{\Sigma}_\zeta\big)\boldsymbol{A}^T + \sigma_\xi^2. \tag{2.36}$$

The addition of the input noise ζ in combination with the unaltered definition of the past information $I(X_0;\boldsymbol{L}_p)$ prohibits solving the Gaussian information bottleneck analytically as outlined above. Based on the problem setup however, we can already make a few observations. Firstly, the joint process of the degraded past input signal $\boldsymbol{S}_p$ and the true future signal ℓ_τ is no longer Markovian, because signal values further in the past can help average out the input noise and therefore affect the conditional probability $p(\ell_\tau|\boldsymbol{S}_p) \neq p(\ell_\tau|s_0)$. Moreover, the input noise can never add information about the future signal. Therefore, the upper bound on the predictive information becomes $I(\boldsymbol{S}_p;\ell_\tau) \leq I(\boldsymbol{L}_p;\ell_\tau) = I(\ell_0;\ell_\tau)$. Secondly, as opposed to the eigenvector solution discussed above, here the compression level of the past input as set by the Lagrangian parameter γ (Eq. 2.10) will affect the optimal kernel shape. This can be understood intuitively when we consider that γ effectively sets the kernel magnitude relative to the intrinsic noise strength σ_ξ^2. For small γ, i.e. low compression, the intrinsic noise magnitude is negligible relative to the kernel amplitude, which amplifies not only the signal but also the input noise. In this regime the system must therefore time average the degraded input signal, to reduce the contribution of the higher-frequency noise. Conversely, for large gamma, corresponding to high compression, the intrinsic noise dominates. In this regime time-averaging is less important, and the kernel magnitude must be raised as much as possible to lift the signal above the intrinsic noise [15, 39].

To determine the optimal mapping $\boldsymbol{A}$ as a function of the compression level γ we could maximize the objective function (Eq. 2.10) numerically for each γ [39]. However, for the purpose of this work the effect of the compression γ is less important. Instead, we are more interested in the general form of the optimal kernel as a function of the signal and noise parameters. Therefore, we here study the optimal kernel in the limit of no compression, $\gamma \to 0$, which allows for an analytical derivation of the optimal integration kernel. In this regime, maximizing the objective function (Eq. 2.10) reduces to maximizing the predictive information.

For Gaussian input signals, maximizing the predictive information is equivalent to minimizing the mean squared prediction error, as can be done by constructing the causal Wiener filter of the input [7, 44]. Earlier work has studied the optimally predictive kernel as predicted by Wiener's filtering theory for an input signal degraded by delta-correlated (white) input noise [7]. Here we extend their derivation to an input signal degraded by colored noise with relative strength σ_ζ and correlation time τ_ζ, as defined in Eq. 2.32.

A brief derivation of the causal Wiener filter is presented in Appendix 2.A. There, the optimal kernel is derived in the frequency domain, such that the resulting expression is the Fourier transform of the optimal integration kernel $k(t)$,

$$K^{\text{opt}}(\omega) = \frac{1}{m_+(\omega)} \left[\frac{S_{s\ell}(\omega)}{m_-(\omega)} e^{i\omega\tau} \right]_+ . \tag{2.37}$$

Throughout this work we use as convention for the Fourier transform and its inverse,

$$F(\omega) = \mathcal{F}\{f(t)\} = \int_{-\infty}^{\infty} dt f(t) e^{-i\omega t}, \tag{2.38}$$

$$f(t) = \mathcal{F}^{-1}\{F(\omega)\} = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega F(\omega) e^{i\omega t}. \tag{2.39}$$

In Eq. 2.37, $S_{s\ell}(\omega)$ is the cross-power spectrum between the degraded input s (Eq. 2.32) and the future signal ℓ. The Wiener-Khinchin theorem tells us that the power spectrum (power spectral density) of a stochastic process is the Fourier transform of its correlation function, such that $S_{s\ell}(\omega) = \mathcal{F}\{\langle \delta s(t')\delta\ell(t)\rangle\}$ with $t' \leq t$. The terms $m_+(\omega)$ and $m_-(\omega)$ arise from the Wiener-Hopf factorization of the power spectrum of the input signal,

$$m_+(\omega)m_-(\omega) = S_s(\omega) = \mathcal{F}\{\langle \delta s(t')\delta s(t'')\rangle\}, \tag{2.40}$$

also see Appendix 2.A. Finally, $[F(\omega)]_+ = \mathcal{F}\{\theta(t)f(t)\}$ with the unit step function θ, defines the causal part of $F(\omega)$.

We find the following factorization for the power spectrum of the signal defined in Eq. 2.32,

$$S_s(\omega) = \frac{2\lambda}{\lambda^2 + \omega^2} + \frac{2\mu\sigma_\zeta^2}{\mu^2 + \omega^2} \tag{2.41}$$

$$= m_+(\omega) \times m_-(\omega) = \frac{\kappa + i\rho\omega}{(\lambda + i\omega)(\mu + i\omega)} \times \frac{\kappa - i\rho\omega}{(\lambda - i\omega)(\mu - i\omega)}, \tag{2.42}$$

with the signal decay rate $\lambda = \tau_\ell^{-1}$ and the noise decay rate $\mu = \tau_\zeta^{-1}$, and where we have set $\sigma_\ell^2 = 1$. We have further defined

$$\kappa = \sqrt{2\lambda\mu(\mu + \lambda\sigma_\zeta^2)}, \tag{2.43}$$

$$\rho = \sqrt{2(\lambda + \mu\sigma_\zeta^2)}. \tag{2.44}$$

See Appendix 3.C for more details on the derivation of power spectra of Gaussian processes starting from their dynamics in Langevin form. The cross-spectrum between the past input and future signal is simply given by the power spectrum of the true signal, because the signal is independent of the input noise,

$$S_{s\ell}(\omega) = S_\ell(\omega) = \frac{2\lambda}{\lambda^2 + \omega^2}. \tag{2.45}$$

Using Eqs. 2.42 and 2.45 in combination with Eq. 2.37 we obtain for the optimal integration kernel in the frequency domain,

$$K^{\mathrm{opt}}(\omega) = \frac{(\lambda + i\omega)(\mu + i\omega)}{\kappa + i\rho\omega} \left[\frac{(\mu - i\omega)2\lambda}{(\kappa - i\rho\omega)(\lambda + i\omega)} e^{i\omega\tau} \right]_+, \tag{2.46}$$

$$= \frac{(\mu + i\omega)(\lambda + \mu)}{(\kappa + i\rho\omega)(\kappa + \lambda\rho)} 2\lambda e^{-\lambda\tau}. \tag{2.47}$$

Finally, taking the inverse Fourier transform yields the optimal integration kernel in the time-domain that maximizes the predictive information,

$$k^{\mathrm{opt}} \propto \delta(t) + (\mu - \kappa/\rho) e^{-t\kappa/\rho}, \tag{2.48}$$

with κ and ρ as defined in Eqs. 2.43 and 2.44 respectively, and where we have omitted the prefactor. The prefactor affects the response amplitude and determines the degree to which the input signal (and input noise) are lifted above the intrinsic noise downstream; in the limit of infinite past information considered here, the added noise downstream is zero, making the prefactor irrelevant.

Equation 2.48 reveals that in general, the optimal kernel is a sum of a delta function with its peak at the most recent signal value $t = 0$, and an exponential tail that averages out the input noise. Its inverse exponent ρ/κ sets the optimal integration time,

$$\tau_{\mathrm{r}}^{\mathrm{opt}} \equiv \rho/\kappa = \tau_\ell \sqrt{1 - \frac{1 - \tau_\zeta/\tau_\ell}{1 + \sigma_\zeta^2 \tau_\zeta/\tau_\ell}}, \tag{2.49}$$

with the signal correlation time $\tau_\ell = \lambda^{-1}$, the noise correlation time $\tau_\zeta = \mu^{-1}$, and the relative noise strength σ_ζ^2. Equation 2.49 shows that the optimal integration time $\tau_{\mathrm{r}}^{\mathrm{opt}}$ increases towards τ_ℓ for both an increasing relative noise strength σ_ζ^2, and an increasing noise correlation time τ_ζ (Fig. 2.1A and B). However, at the same time, the weight given to the exponential tail relative to the peak at $t = 0$ decreases with the noise correlation time (Eq. 2.48 and Fig. 2.1A). While this might seem counter-intuitive, the reason is that noise with a longer correlation time is increasingly difficult to average out. Indeed, when the correlation time of the noise equals that of the signal, $\tau_\zeta = \tau_\ell$, time-averaging is no longer feasible because the kernel would also average out the signal. In this regime the optimal kernel reduces to a delta function only (Eq. 2.48), as for a Markovian signal without input noise. In the other limit, when the correlation time of the noise becomes very short, the weight of the exponential tail increases relative to the delta function while its length, set by $\tau_{\mathrm{r}}^{\mathrm{opt}}$, decreases (Eq. 2.49, Fig. 2.1A). We can take the limit of white noise

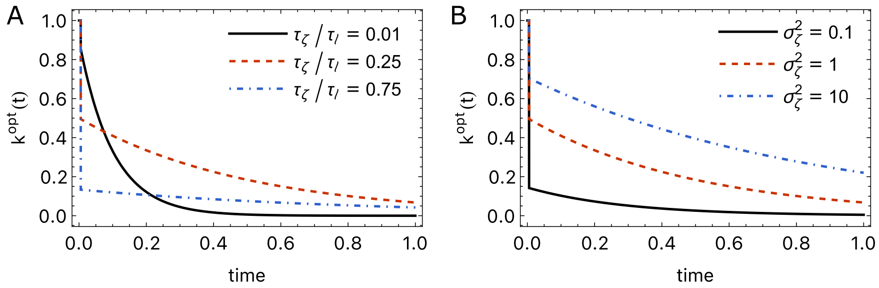


Figure 2.1: The optimal integration kernel in the limit $\gamma \to 0$ for an input signal degraded by colored noise. (A) The optimal kernel for different noise correlation times τ_ζ, relative noise strength $\sigma_\zeta^2 = 1$. (B) The optimal kernel for different noise strengths σ_ζ^2, noise correlation time $\tau_\zeta/\tau_\ell = 0.25$.

by defining a white noise strength that is kept constant $\vartheta^2 = 2\sigma_\zeta^2\tau_\zeta$ while taking the limit $\tau_\zeta \to 0$. We then obtain,

$$\lim_{\tau_\zeta \to 0} k^{\text{opt}} \propto \exp(-t\sqrt{\lambda^2 + 2\lambda/\vartheta^2}), \tag{2.50}$$

which agrees with the optimal kernel of Becker *et al.* for the prediction of an input signal degraded by white noise [7].

2.5. NON-MARKOVIAN INPUT

Not all ligand concentration trajectories encountered by cells are expected to be Markovian. For example, the bacterium *Escherichia coli* swims in its environment with a speed which exhibits persistence. This leads to an auto-correlation function for the concentrations' derivative which does not decay instantaneously [20]. To model such a persistent signal, we use the classical model of a particle in a harmonic well or a mass on a spring, also known as a stochastically driven damped harmonic oscillator

$$\begin{aligned} \delta\dot{\ell} &= v(t), \\ \dot{v} &= -\omega_0^2\delta\ell(t) - v(t)/\tau_v + \eta_v(t), \end{aligned} \tag{2.51}$$

where $\omega_0 = \sqrt{k/m}$, with k the spring constant and m the mass of the particle, τ_v is the relaxation time, and $\eta_v(t) = \sigma_v\sqrt{2/\tau_v}\xi(t)$, with $\xi(t)$, as used throughout, a Gaussian white noise process of unit variance, and σ_v the standard deviation of v. If the signal would obey the fluctuation-dissipation relation, then $m\sigma_v^2 = k_\text{B}T$, but since the biochemical signal could very well be generated via an active process this relation may not hold.

We can summarize Eq. 2.51 as a 2-dimensional OU-process,

$$\dot{\boldsymbol{s}} = \boldsymbol{J}\boldsymbol{s}(t) + \boldsymbol{B}\boldsymbol{\xi}(t) \tag{2.52}$$

where $\boldsymbol{s}(t) = (\delta\ell(t), v(t))^T$ is the signal, $\boldsymbol{\xi}(t)$ is a 2×1 vector of unit white noise processes, and the Jacobian and noise matrix of the signal respectively are

$$\boldsymbol{J} = \begin{pmatrix} 0 & 1 \\ -\omega_0^2 & -1/\tau_v \end{pmatrix}, \tag{2.53}$$

$$\boldsymbol{B} = \begin{pmatrix} 0 & 0 \\ 0 & \sigma_v\sqrt{2/\tau_v} \end{pmatrix}. \tag{2.54}$$

This way of defining the signal dynamics (Eq. 2.52) allows us to directly exploit the Lyapunov equation,

$$\boldsymbol{JC} + \boldsymbol{CJ}^T + \boldsymbol{BB}^T = \boldsymbol{0}, \tag{2.55}$$

where $\boldsymbol{C}$ is the covariance matrix of the signal. The correlation matrix of the signal is then given by

$$\boldsymbol{C}(\tau) = e^{\boldsymbol{J}\tau}\boldsymbol{C} \qquad \text{for } \tau > 0. \tag{2.56}$$

For a derivation of this result, see for example the work by Vennettilli *et al.* [45]. In the chapters that follow we will exploit the power spectra of stochastic processes to compute the required statistics, which is explained in detail in Appendix 3.C. For briefness, we here simply use Eqs. 2.55 and 2.56 to derive the covariance matrix and correlation functions of the signal.

We first solve Eq. 2.55 for the covariance matrix $\boldsymbol{C}$ and obtain

$$\boldsymbol{C} = \begin{pmatrix} \sigma_\ell^2 & \sigma_{\ell v} \\ \sigma_{\ell v} & \sigma_v^2 \end{pmatrix} = \sigma_v^2 \begin{pmatrix} 1/\omega_0^2 & 0 \\ 0 & 1 \end{pmatrix}. \tag{2.57}$$

Using Eq. 2.56 we then obtain the auto-correlation matrix in the underdamped regime, $\tau_v^{-1} < 2\omega_0$,

$$\begin{aligned}\boldsymbol{C}(\tau) &= \begin{pmatrix} \langle\delta\ell(\tau)\delta\ell(0)\rangle & \langle\delta\ell(\tau)\delta v(0)\rangle \\ \langle\delta v(\tau)\delta\ell(0)\rangle & \langle\delta v(\tau)\delta v(0)\rangle \end{pmatrix}, \\ &= \begin{pmatrix} \sigma_\ell^2 e^{-\mu\tau/2}\left(\cos(\rho'\tau) + \frac{\mu}{2\rho'}\sin(\rho'\tau)\right) & \sigma_v^2 e^{-\mu\tau/2}\frac{1}{\rho'}\sin(\rho'\tau) \\ -\sigma_v^2 e^{-\mu\tau/2}\frac{1}{\rho'}\sin(\rho'\tau) & \sigma_v^2 e^{-\mu\tau/2}\left(\cos(\rho'\tau) - \frac{\mu}{2\rho'}\sin(\rho'\tau)\right) \end{pmatrix}.\end{aligned} \tag{2.58}$$

where $\rho' = \sqrt{\omega_0^2 - \mu^2/4}$, with $\mu = \tau_v^{-1}$. We can readily obtain the correlation functions in the overdamped regime, $\tau_v^{-1} > 2\omega_0$, from Eq. 2.58 using Euler's formula's,

$$e^{ix} = \cos x + i\sin x, \tag{2.59}$$

$$e^{-ix} = \cos x - i\sin x, \tag{2.60}$$

which yield $\sin x = \left(e^{ix} - e^{-ix}\right)/2i$ and $\cos x = \left(e^{ix} + e^{-ix}\right)/2$. To relate sine and cosine to their hyperbolic counterpart we use the definitions of the hyperbolic functions,

$$\sinh x = \left(e^x - e^{-x}\right)/2 = -i\sin(ix), \tag{2.61}$$

$$\cosh x = \left(e^x - e^{-x}\right)/2 = \cos(ix). \tag{2.62}$$

We define $\rho = \sqrt{\mu^2/4 - \omega_0^2} = i\rho'$, which also means that $\rho' = i\rho$, such that $\cos(\rho'\tau) = \cosh(\rho\tau)$ and $\sin(\rho'\tau) = i\sinh(\rho\tau)$. Substituting $\rho' = i\rho$ in Eq. 2.58 then yields the correlation functions in the overdamped regime, $\mu = \tau_v^{-1} > 2\omega_0$,

$$\boldsymbol{C}(\tau) = \begin{pmatrix} \sigma_\ell^2 e^{-\mu\tau/2}\left(\cosh(\rho\tau) + \frac{\mu}{2\rho}\sinh(\rho\tau)\right) & \sigma_v^2 e^{-\mu\tau/2}\frac{1}{\rho}\sinh(\rho\tau) \\ -\sigma_v^2 e^{-\mu\tau/2}\frac{1}{\rho}\sinh(\rho\tau) & \sigma_v^2 e^{-\mu\tau/2}\left(\cosh(\rho\tau) - \frac{\mu}{2\rho}\sinh(\rho\tau)\right) \end{pmatrix}. \tag{2.63}$$

Using the correlation functions shown above we can now derive the optimal mapping of the past ligand trajectory onto the current output that enables the sensing system to reach the information bound. In general, the optimal mapping depends on whether the system needs to predict the signal concentration or the signal derivative in the future, or both. Here we will first focus on the prediction of the derivative, because this is arguably more informative than the absolute concentration for cells navigating chemical gradients. However, it is straightforward to apply the same analysis for prediction of the future signal concentration. In fact, we will find that regardless of the prediction objective the optimal mapping is determined by the correlation between the current signal concentration and the future signal property of interest, and by the correlation between the current signal derivative and the future signal property of interest. In Section 2.6 we return to the optimal prediction of both the signal concentration and its derivative, which requires two output components given sufficiently high past information.

To find the optimal mapping for the prediction of the derivative of a non-Markovian signal based on its past ligand concentration trajectory, we need to find the eigenvalues and vectors of the matrix (see Eqs. 2.16 and 2.17)

$$\begin{aligned} \boldsymbol{W} &= \boldsymbol{\Sigma}_{L|v_\tau}\boldsymbol{\Sigma}_L^{-1}, \\ &= \mathbb{I}_N - \frac{1}{\sigma_v^2}\boldsymbol{\Sigma}_{Lv_\tau}\boldsymbol{\Sigma}_{Lv_\tau}^T\boldsymbol{\Sigma}_L^{-1}, \end{aligned} \tag{2.64}$$

where we again used the Schur complement formula (Eq. 2.13) to rewrite the conditional covariance matrix in terms of the cross-correlation matrices. The covariance matrix of the past trajectory is symmetric with entries $\boldsymbol{\Sigma}_L^{(i,j)} = \langle\delta\ell(0)\delta\ell(|i-j|\Delta t)\rangle$ where both i and j range from 1 to N, the past trajectory length. The covariance vector between past trajectory and future derivative has entries $\boldsymbol{\Sigma}_{Lv_\tau}^{(i,j)} = \langle\delta\ell(0)\delta v(\tau + (i-1)\Delta t)\rangle$. Both the concentration auto-correlation function and the concentration to future derivative cross-correlation function are shown in Eqs. 2.58 and 2.63, depending on whether the system is overdamped ($\tau_v^{-1} > 2\omega_0$) or underdamped ($\tau_v^{-1} < 2\omega_0$).

To better understand the optimal mapping of this signal we numerically investigate the eigenvalues of the matrix $\boldsymbol{W}$. For the prediction of the future derivative v_τ, there is only one non-trivial eigenvalue. Like for the Markovian signal, this shows that for the prediction of the derivative of this non-Markovian signal, the optimal mapping is always onto a scalar output. The non-trivial eigenvalue λ decreases with the discretization timestep Δt and is minimal for $\Delta t \to 0$ (Fig. 2.2). In this limit, λ has the same magnitude for any $N \geq 2$, see Fig. 2.2. A smaller eigenvalue λ corresponds to larger past and predictive information and a larger ratio $I_{\text{pred}}/I_{\text{past}}$ (Eq. 2.19 and Eq. 2.20), given any value

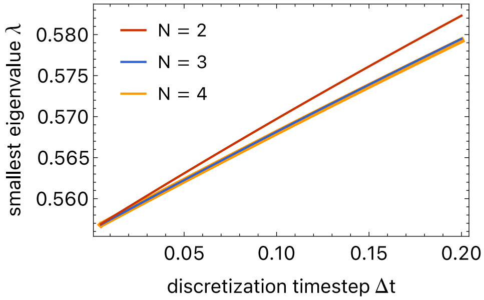


Figure 2.2: The smallest eigenvalue of the IB matrix is minimal for $N \geq 2$ and $\Delta t \to 0$. A smaller eigenvalue corresponds to a larger ratio $I_{\text{pred}}/I_{\text{past}}$ for any given value of the Lagrange multiplier γ. Parameters: friction timescale $\tau_v^{-1} = 0.862s^{-1}$ as determined in [20], prediction interval $\tau = \tau_v$, and $\omega_0 = 0.4s^{-1}$. While the magnitude of the eigenvalues does depend on the parameter values, the qualitative picture remains the same, i.e. the smallest eigenvalue of the IB matrix remains minimal for $N \geq 2$ and $\Delta t \to 0$ regardless of the choice of parameters.

of the Lagrange multiplier γ. For the optimal mapping we must thus have $N \geq 2$ and $\Delta t \to 0$, where N sets both the past trajectory and the mapping vector length. Because increasing the length above two does not yield an improvement in the value of λ_1 we focus on $N = 2$.

The fact that to reach the optimum we must have $N = 2$ and $\Delta t \to 0$, shows that the optimal kernel $\boldsymbol{A}$ takes an instantaneous measurement of a combination of the most recent ligand concentration, and its derivative. This can be understood by noting that for a trajectory of length two, the mapping vector also has two components, $\boldsymbol{A} = ||A||(\hat{w}_1, \hat{w}_2)$, with $\sqrt{\hat{w}_1^2 + \hat{w}_2^2} = 1$. We can then express the linear mapping of $\boldsymbol{L}_p$ to X_0 (Eq. 2.11) as:

$$X_0 = ||A|| \left[(\hat{w}_1 + \hat{w}_2)\delta\ell(0) - \hat{w}_2 \Delta t \frac{\delta\ell(0) - \delta\ell(-\Delta t)}{\Delta t} \right] + \xi, \tag{2.65}$$

This expression shows that, as $\Delta t \to 0$, the two entries of $\boldsymbol{A}$ combine both the most recent signal value and the most recent derivative to generate X_0. This is intuitive because the signal is completely defined by its concentration and derivative (Eq. 2.51). For this reason, and to obtain analytical insight into the optimal weights, we inspect the final two entries of the past ligand concentration trajectory in the limit $\Delta t \to 0$, which defines the past signal in terms of its most recent concentration and derivative

$$\boldsymbol{S}_0 \equiv \begin{pmatrix} \delta\ell(0) & v(0) \end{pmatrix}^T. \tag{2.66}$$

Because the signal is Markovian in the joint properties $\delta\ell$ and v, the vector $\boldsymbol{S}_0$ contains the same information as the trajectory $\boldsymbol{L}_p$. The past information is now the mutual information between X_0 and $\boldsymbol{S}_0$, i.e. $I_{\text{past}} = I(X_0; \boldsymbol{S}_0)$. The output X_0 can then also be written as a projection of $\boldsymbol{S}_0$ via the alternative mapping vector $\tilde{\boldsymbol{A}} = ||A||(\hat{a}, \hat{b})$:

$$X_0 = ||A|| \left(\hat{a}\delta\ell(0) + \hat{b}v(0) \right) + \xi. \tag{2.67}$$

Comparison with Eq. 2.65 shows how the components of $\tilde{\boldsymbol{A}}$ relate back to those of $\boldsymbol{A}$,

$$\hat{w}_1 = \hat{a} + \hat{b}/\Delta t, \tag{2.68}$$

$$\hat{w}_2 = -\hat{b}/\Delta t. \tag{2.69}$$

To obtain the optimal mapping vector $\tilde{\boldsymbol{A}}$ the matrix of signal statistics of which the eigenvalues and -vectors need to be determined is

$$\boldsymbol{W} = \boldsymbol{\Sigma}_{\boldsymbol{s}|v_\tau}\boldsymbol{\Sigma}_{\boldsymbol{s}}^{-1}, \tag{2.70}$$

with

$$\boldsymbol{\Sigma}_{\boldsymbol{s}} = \begin{pmatrix} \sigma_\ell^2 & 0 \\ 0 & \sigma_v^2 \end{pmatrix}, \tag{2.71}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{s}|v_\tau} = \boldsymbol{\Sigma}_{\boldsymbol{s}} - \frac{1}{\sigma_v^2}\boldsymbol{\Sigma}_{\boldsymbol{s}v_\tau}\boldsymbol{\Sigma}_{\boldsymbol{s}v_\tau}^T, \tag{2.72}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{s}v_\tau} = \begin{pmatrix} \langle\delta\ell(0)\delta v(\tau)\rangle \\ \langle\delta v(0)\delta v(\tau)\rangle \end{pmatrix}. \tag{2.73}$$

We thus obtain

$$\boldsymbol{W} = \mathbb{I} - \begin{pmatrix} \frac{\langle\delta\ell(0)\delta v(\tau)\rangle^2}{\sigma_\ell^2\sigma_v^2} & \frac{\langle\delta\ell(0)\delta v(\tau)\rangle\langle\delta v(0)\delta v(\tau)\rangle}{\sigma_v^4} \\ \frac{\langle\delta\ell(0)\delta v(\tau)\rangle\langle\delta v(0)\delta v(\tau)\rangle}{\sigma_\ell^2\sigma_v^2} & \frac{\langle\delta v(0)\delta v(\tau)\rangle^2}{\sigma_v^4} \end{pmatrix}. \tag{2.74}$$

This matrix has one nontrivial eigenvalue, $\lambda = 1 - \frac{\langle\delta v(0)\delta v(\tau)\rangle^2}{\sigma_v^4} - \frac{\langle\delta\ell(0)\delta v(\tau)\rangle^2}{\sigma_\ell^2\sigma_v^2}$, which depends on the normalized correlation functions between on the one hand the current concentration or derivative, and on the other hand the future derivative. The corresponding left eigenvector is

$$\boldsymbol{v}_1 = Q^{-1}\begin{pmatrix} \frac{1}{\sigma_\ell}\frac{\langle\delta\ell(0)\delta v(\tau)\rangle}{\sigma_\ell\sigma_v} & \frac{1}{\sigma_v}\frac{\langle\delta v(0)\delta v(\tau)\rangle}{\sigma_v^2} \end{pmatrix}, \tag{2.75}$$

where Q normalizes the vector. Using the linear mapping $X_0 = ||A||\boldsymbol{v}_1\boldsymbol{S}_0 + \xi$, and defining $G \equiv ||A||/Q$, shows that the optimal output should be generated as follows

$$X_0^{\text{opt}} = G\left(\frac{\langle\delta\ell(0)\delta v(\tau)\rangle}{\sigma_\ell\sigma_v}\frac{\delta\ell(0)}{\sigma_\ell} + \frac{\langle\delta v(0)\delta v(\tau)\rangle}{\sigma_v^2}\frac{v(0)}{\sigma_v}\right) + \xi. \tag{2.76}$$

We thus find that the optimal mapping depends on the (normalized) cross-correlation coefficient $\rho_{\ell_0 v_\tau} \equiv \langle\delta\ell(0)\delta v(\tau)\rangle/(\sigma_\ell\sigma_v)$ between the current concentration $\delta\ell(0)$ and future derivative $\delta v(\tau)$, and the cross-correlation coefficient $\rho_{v_0 v_\tau}$ between the current derivative $\delta v(0)$ and future derivative $\delta v(\tau)$. Indeed, to optimally predict the future derivative, the cell should also use the current concentration and not only the current derivative. The reason is that for an underdamped signal with inertia not only the current derivative but also the current value is correlated with, and hence informative about, the future derivative. Repeating the analysis above (Eqs. 2.70-2.76) for the prediction of the future signal concentration instead of the future derivative shows that generally, the cell should use both the current signal concentration and derivative and weigh them relative to their correlation with the future signal property of interest.

2.6. PREDICTING MULTIPLE COMPONENTS

Up to this point, we have considered systems that predict either the concentration or the derivative of a given input signal. In those cases, it turned out that the optimal mapping can always be made onto a scalar output. Here, we will consider what happens when a system aims to predict multiple signals or multiple signal features, e.g. both the concentration and the derivative of a signal with the dynamics given in Eq. 2.51.

In their work on the Gaussian information bottleneck, Chechik *et al.* [16] show that in such a scenario, under high compression, the optimal mapping remains onto a scalar output X. However, when the compression decreases, such that the system can gather more past information, an additional mapping component emerges, and the optimal mapping is onto a two-dimensional output $\boldsymbol{X}$. The optimal mapping for each component is given by the corresponding eigenvector of a matrix defined by the signal statistics, as also derived above for a univariate mapping (Eq. 2.18) [16]. On the one hand, it is perhaps not surprising that a system may require multiple independent output components to optimally predict multiple signals, or multiple signal features. On the other hand, the result that a system should use only one output component under high compression, and add more output components as the compression level decreases past given transition points is highly intriguing. Why is this strategy preferable to that of always using multiple components, thus even in the regime of high compression, using smaller amplitudes for the different output components? And what is the intuition behind the mathematically well-defined transition points? These are the questions we aim to answer in this section.

To gain a simpler and perhaps more intuitive picture of the Gaussian information bottleneck we consider a signal characterized by two distinct features $\boldsymbol{S}(t) = (\delta\ell(t), v(t))^T$ (also see Section 2.5) which is normalized to ensure that its covariance matrix is the identity matrix $\Sigma_s = \mathbb{I}_2$. In Appendix 2.B we demonstrate how such a normalization can be performed for any multivariate input signal, and how the resulting optimal kernels may be mapped back onto the original signal basis. Therefore, this approach retains complete generality for any input signal statistics.

Given that the signal covariance matrix is the identity matrix, the optimal kernels for prediction are now given by the left eigenvectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ of the matrix $\Sigma_{\boldsymbol{s}_0|\boldsymbol{s}_\tau}\Sigma_{\boldsymbol{s}_0}^{-1} = \Sigma_{\boldsymbol{s}_0|\boldsymbol{s}_\tau}$ corresponding to the eigenvalues λ_1 and λ_2, in order of increasing magnitude (see [16] or Eq. 2.16). Moreover, in what follows, instead of normalizing the noise in the output and optimizing the amplitude $||A||$ of the mapping from signal to output, we consider the mapping amplitude to be normalized and optimize the noise magnitude. We thus have for the two-dimensional output of a linear signaling system,

$$\boldsymbol{X}(t) = \boldsymbol{N}\boldsymbol{S}(t) + \boldsymbol{\xi}(t), \tag{2.77}$$

where $\boldsymbol{X}(t) = (x_1(t), x_2(t))^T$ and

$$\boldsymbol{N} = \begin{pmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \end{pmatrix}, \tag{2.78}$$

with $||\boldsymbol{v}_1|| = 1$ and $||\boldsymbol{v}_2|| = 1$. The covariance matrix of the white noise vector $\boldsymbol{\xi}$ is

$$\boldsymbol{\Sigma}_\xi = \begin{pmatrix} \sigma^2_{\xi_1} & 0 \\ 0 & \sigma^2_{\xi_2}. \end{pmatrix} \tag{2.79}$$

The change from considering normalized noise, as in the sections above, to considering a normalized kernel here, is made purely for simplicity of the mathematical expressions. Following the argument that was also made by Chechik *et al.* [16], for any mapping and noise covariance pair $(\boldsymbol{A}, \boldsymbol{\Sigma}_\xi)$, we can consider a normalized mapping and rescaled noise $(\boldsymbol{N}, \boldsymbol{\Sigma}'_\xi)$ that yield the same past and predictive information.

We consider the past and predictive information from a 'decoding' perspective, i.e. asking to what extent knowledge of the current output $\boldsymbol{X}_0$ reduces the uncertainty in the past or future signal, $\boldsymbol{S}_0$ and $\boldsymbol{S}_\tau$ respectively. Using the definition of the mutual information between Gaussian random variables (Eq. 2.7) we find

$$I_{\text{past}}(\boldsymbol{S}_0; \boldsymbol{X}_0) = \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_{\boldsymbol{s}_0}|}{|\boldsymbol{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{x}_0}|} = -\frac{1}{2} \log |\boldsymbol{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{x}_0}|, \tag{2.80}$$

$$I_{\text{pred}}(\boldsymbol{S}_\tau; \boldsymbol{X}_0) = \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_{\boldsymbol{s}_\tau}|}{|\boldsymbol{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{x}_0}|} = -\frac{1}{2} \log |\boldsymbol{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{x}_0}|, \tag{2.81}$$

where we used that $\boldsymbol{\Sigma}_{\boldsymbol{s}_0} = \boldsymbol{\Sigma}_{\boldsymbol{s}_\tau} = \mathbb{I}$. The determinants of the covariance matrices quantify the spread of the corresponding distribution. Equation 2.81 shows that the predictive information increases as the distribution $p(\boldsymbol{S}_\tau|\boldsymbol{X}_0)$ becomes more sharply peaked and the determinant of the corresponding covariance matrix thus shrinks, i.e. the uncertainty about the future signal given the current system output becomes smaller. The same reasoning holds for the past information. Note that, because the marginal distributions are normalized such that their covariance matrix' determinants are unity, the determinants of the conditional covariance matrices are always smaller than one.

PAST INFORMATION

To quantify the past information we require the determinant of the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{x}_0}$ (Eq. 2.80). Using the Schur complement formula (Eq. 2.13), we can express this matrix as

$$\boldsymbol{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{x}_0} = \mathbb{I} - \boldsymbol{\Sigma}_{\boldsymbol{sx}} \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{sx}}^T, \tag{2.82}$$

where via the definition of $\boldsymbol{X}$ (Eq. 2.77) we have $\boldsymbol{\Sigma}_{\boldsymbol{sx}} = \mathbb{E}\left[\boldsymbol{S}\boldsymbol{X}^T\right] = \boldsymbol{N}^T$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{N}\boldsymbol{N}^T + \boldsymbol{\Sigma}_\xi$. Moreover, the product of mapping matrices $\boldsymbol{N}\boldsymbol{N}^T$ simplifies greatly when we consider that its components are left eigenvectors of the matrix $\boldsymbol{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{s}_\tau}$, which are by definition orthogonal,

$$\boldsymbol{N}\boldsymbol{N}^T = \begin{pmatrix} \boldsymbol{v}_1\boldsymbol{v}_1^T & \boldsymbol{v}_1\boldsymbol{v}_2^T \\ \boldsymbol{v}_2\boldsymbol{v}_1^T & \boldsymbol{v}_2\boldsymbol{v}_2^T \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{2.83}$$

The covariance in the current signal given the current output thus becomes,

$$\boldsymbol{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{x}_0} = \mathbb{I} - \boldsymbol{N}^T \left(\mathbb{I} + \boldsymbol{\Sigma}_\xi\right)^{-1} \boldsymbol{N}, \tag{2.84}$$

$$= \mathbb{I} - \frac{\boldsymbol{v}_1^T \boldsymbol{v}_1}{1 + \sigma^2_{\xi_1}} - \frac{\boldsymbol{v}_2^T \boldsymbol{v}_2}{1 + \sigma^2_{\xi_2}}. \tag{2.85}$$

Conveniently, this expression reveals that $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are also left eigenvectors of the matrix $\boldsymbol{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{x}_0}$, since

$$\boldsymbol{v}_1\boldsymbol{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{x}_0} = \boldsymbol{v}_1\sigma^2_{\xi_1}/(1+\sigma^2_{\xi_1}), \tag{2.86}$$

$$\boldsymbol{v}_2\boldsymbol{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{x}_0} = \boldsymbol{v}_2\sigma^2_{\xi_2}/(1+\sigma^2_{\xi_2}). \tag{2.87}$$

The corresponding eigenvalues are thus

$$\alpha_1 = \sigma^2_{\xi_1}/(1+\sigma^2_{\xi_1}), \tag{2.88}$$

$$\alpha_2 = \sigma^2_{\xi_2}/(1+\sigma^2_{\xi_2}). \tag{2.89}$$

These eigenvectors and eigenvalues quantify the principal directions and magnitude of the uncertainty in the current signal $\boldsymbol{S}_0$ given the current output $\boldsymbol{X}_0$, as illustrated in the left panel of Fig. 2.3B. The maximal uncertainty in each direction is one, and occurs when the noise in the corresponding mapping components diverges, $\sigma^2_{\xi_i} \to \infty$. The minimal uncertainty in each direction is zero, when the corresponding noise term vanishes. This observation agrees with intuition: when there is no noise in the mapping, the system output is a perfect copy of the signal and there is no uncertainty in the current signal value given the current output.

Finally, the determinant of a matrix is given by the product of its eigenvalues and we therefore obtain for the past information (Eq. 2.80)

$$I_{\text{past}}(\boldsymbol{S}_0;\boldsymbol{X}_0) = -\frac{1}{2}\log(\alpha_1\alpha_2) = \frac{1}{2}\log(1+1/\sigma^2_{\xi_1}) + \frac{1}{2}\log(1+1/\sigma^2_{\xi_2}), \tag{2.90}$$

where $1/\sigma^2_{\xi_1}$ and $1/\sigma^2_{\xi_2}$ represent the signal to noise ratio in the two mapping directions $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ (compare to Eq. 2.19 where here $||A||^2 = 1$ and $\boldsymbol{\Sigma}_{\boldsymbol{s}} = \mathbb{I}$, while $\sigma^2_{\xi} \neq 1$). When only one component is used in the mapping and the noise of the second component is thus infinitely large, i.e. $\sigma^2_{\xi_2} \to \infty$ and $\alpha_2 = 1$ (Eq. 2.89), then the past information is set by the noise in the first component only (Eq. 2.90).

PREDICTIVE INFORMATION

The predictive information can be determined in a similar way as the past information, i.e. by leveraging the eigenvalues of the conditional covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{x}_0}$. This conditional covariance matrix is given by, via the Schur complement formula (Eq. 2.13),

$$\boldsymbol{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{x}_0} = \mathbb{I} - \boldsymbol{\Sigma}_{\boldsymbol{s}_\tau\boldsymbol{x}}\boldsymbol{\Sigma}^{-1}_{\boldsymbol{x}}\boldsymbol{\Sigma}^T_{\boldsymbol{s}_\tau\boldsymbol{x}}, \tag{2.91}$$

which using $\boldsymbol{\Sigma}_{\boldsymbol{s}_\tau\boldsymbol{x}} = \mathbb{E}\left[\boldsymbol{S}(\tau)\boldsymbol{X}_0^T\right] = \boldsymbol{\Sigma}^T_{\boldsymbol{s}\boldsymbol{s}_\tau}\boldsymbol{N}^T$ and again substituting $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ (using Eqs. 2.77 and 2.83) becomes,

$$\boldsymbol{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{x}_0} = \mathbb{I} - \boldsymbol{\Sigma}^T_{\boldsymbol{s}\boldsymbol{s}_\tau}\boldsymbol{N}^T\left(\mathbb{I}+\boldsymbol{\Sigma}_\xi\right)^{-1}\boldsymbol{N}\boldsymbol{\Sigma}_{\boldsymbol{s}\boldsymbol{s}_\tau}, \tag{2.92}$$

$$= \mathbb{I} - \frac{\boldsymbol{\Sigma}^T_{\boldsymbol{s}\boldsymbol{s}_\tau}\boldsymbol{v}_1^T\boldsymbol{v}_1\boldsymbol{\Sigma}_{\boldsymbol{s}\boldsymbol{s}_\tau}}{1+\sigma^2_{\xi_1}} - \frac{\boldsymbol{\Sigma}^T_{\boldsymbol{s}\boldsymbol{s}_\tau}\boldsymbol{v}_2^T\boldsymbol{v}_2\boldsymbol{\Sigma}_{\boldsymbol{s}\boldsymbol{s}_\tau}}{1+\sigma^2_{\xi_2}}. \tag{2.93}$$

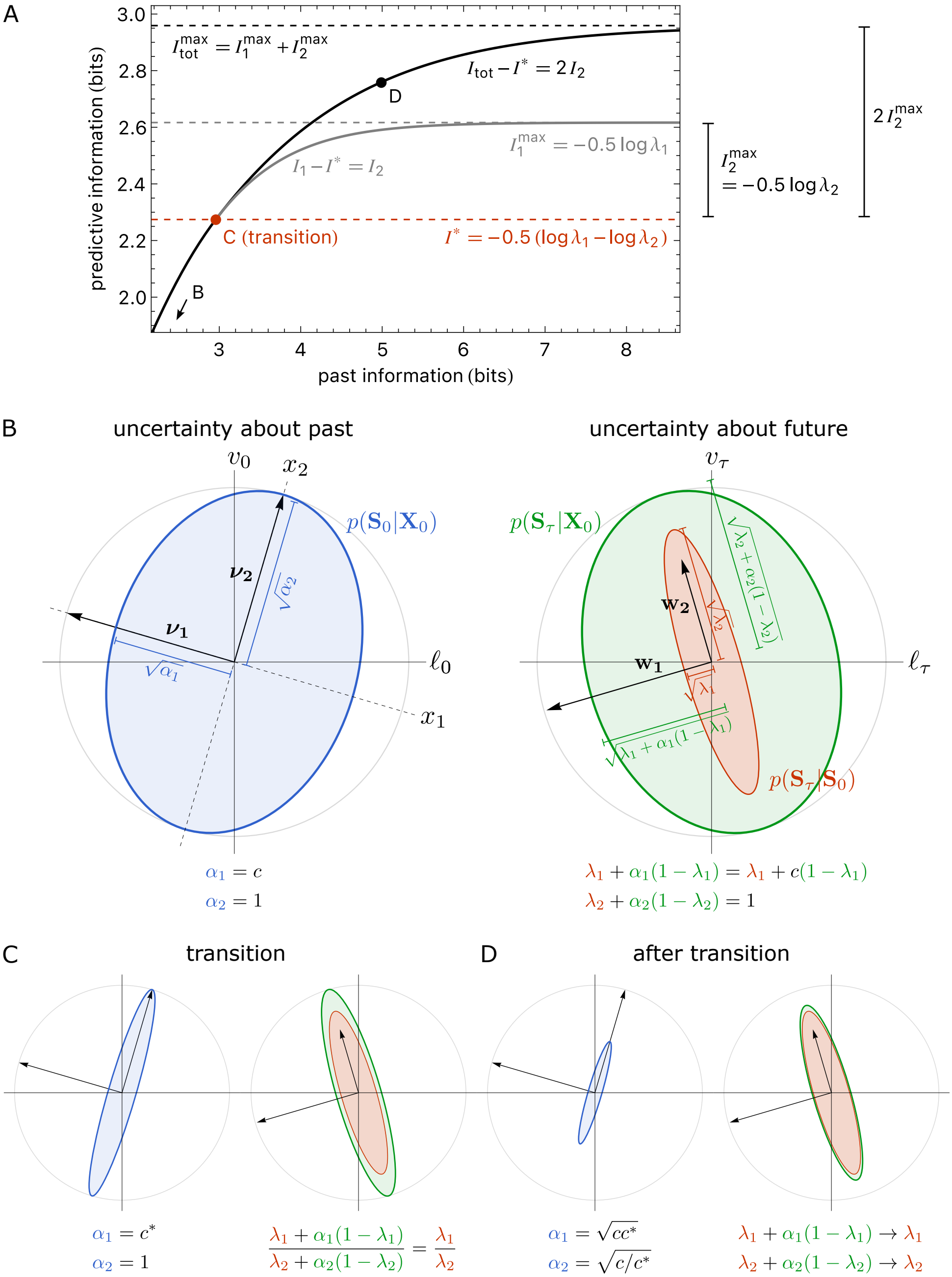


Figure 2.3: The information about the past and future signal in the current output and the optimal number of mapping components are set by the spread in the corresponding distributions. Caption on the following page.

Figure 2.3: Figure on previous page. (A) The information bound (black curve) showing the maximal predictive information as a function of the past information for the signal defined in Eq. 2.51 with $\sigma_v^2 = \sigma_\ell^2 = 1$, $\tau_v^{-1} = 0.862\mathrm{s}^{-1}$, and $\tau = \tau_v/2$. Increasing the past information $I_{\mathrm{past}} = -1/2 \log|\boldsymbol{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{x}_0}|$ corresponds to reducing the uncertainty about the past signal $|\boldsymbol{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{x}_0}| = c$ (Eq. 2.108). Initially, for low past information, the system only uses the output component with the largest encoding capacity, such that $I_{\mathrm{pred}} = I_1 = -1/2\log(\lambda_1 + c(1-\lambda_1))$ (Eq. 2.106 with $\alpha_1 = c$, see also panel B). This is the optimal strategy until the transition where $I_{\mathrm{past}} = -1/2\log c^*$ and $I_{\mathrm{pred}} = I^* = I_1(c^*) = -1/2(\log\lambda_1 - \log\lambda_2)$ (red dashed line, see also Eq. 2.112 and panel C). Beyond this point, both components are used such that $I_{\mathrm{pred}} = I_{\mathrm{tot}} = I_1 + I_2$, with $I_1 = -1/2\log\left(\lambda_1\left[1+\sqrt{c(1-\lambda_1)(1-\lambda_2)/(\lambda_1\lambda_2)}\right]\right)$ and $I_2 = -1/2\log\left(\lambda_2\left[1+\sqrt{c(1-\lambda_1)(1-\lambda_2)/(\lambda_1\lambda_2)}\right]\right)$ (see Eqs. 2.113 and 2.114 and panel D). The information contributed by the first component, I_1, after the transition point is shown in gray. Subtracting from this gray curve the predictive information I^* at the transition point, yields the information curve of the second component only, I_2. (B) Left: The distribution $p(\boldsymbol{S}_0|\boldsymbol{X}_0)$ of the current signal given the current output, visualized via an iso-density ellipse defined by the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{x}_0}$. The area of the ellipse is proportional to the square root of the matrix determinant $|\boldsymbol{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{x}_0}|^{1/2} = \sqrt{\alpha_1\alpha_2}$. The square root of the matrix eigenvalues $\alpha_1 = \sigma_{\xi_1}^2/(1+\sigma_{\xi_1}^2)$ and $\alpha_2 = \sigma_{\xi_2}^2/(1+\sigma_{\xi_2}^2)$ respectively set the length of the semi-minor and semi-major axis of the ellipse, where $\sigma_{\xi_1}^2$ and $\sigma_{\xi_2}^2$ are the noise in each output component (Eq. 2.79). The light gray unit circle represents the marginal signal distribution with $\Sigma_{\boldsymbol{s}} = \mathbb{I}_2$. The eigenvectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ of the matrix $\boldsymbol{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{x}_0}$ are the optimal mapping directions of the input signal $\boldsymbol{S}_0$, which minimize the uncertainty about the future, $|\boldsymbol{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{x}_0}|$ (Eq. 2.109), for given uncertainty about the past, $|\boldsymbol{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{x}_0}|$ (Eq. 2.108). They map the signal onto the output components x_1 and x_2 respectively. Note that the length of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ is unity independent of I_{past} because in our description here we keep the kernel fixed but vary I_{past} by varying the noise. Right: similar to the left panel but for the distribution $p(\boldsymbol{S}_\tau|\boldsymbol{X}_0)$ of the future signal given the current output (outer green ellipse) and the distribution $p(\boldsymbol{S}_\tau|\boldsymbol{S}_0)$ of the future signal given the current signal (inner red ellipse). The ellipse axes lengths are set by the (square roots of) eigenvalues λ_1 and λ_2 of the matrix $\boldsymbol{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{s}_0}$ (inner red), and the (square roots of) eigenvalues $\lambda_1 + \alpha_1(1-\lambda_1)$ and $\lambda_2 + \alpha_2(1-\lambda_2)$ of the matrix $\boldsymbol{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{x}_0}$ (outer green), as indicated in the figure. The eigenvectors $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ of both conditional covariance matrices correspond to a mapping of the vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ onto the basis of the future signal $\boldsymbol{S}_\tau$ (Eqs. 2.99 and 2.100). The lengths of these vectors, $||\boldsymbol{w}_1|| = \sqrt{1-\lambda_1}$ and $||\boldsymbol{w}_2|| = \sqrt{1-\lambda_2}$ (Eq. 2.102), are independent of I_{past}. The constraint on the past uncertainty (in both panels) is $c = \alpha_1\alpha_2 = 0.4$, setting the past information via $I_{\mathrm{past}} = -1/2\log c$ (Eq. 2.90). This value of c is larger than the critical value, $c > c^*$ (Eq. 2.112), corresponding to $I_{\mathrm{past}} < I_{\mathrm{past}}^*$, such that the optimal system uses only the first mapping component $\boldsymbol{v}_1$, i.e. $\sigma_{\xi_2}^2 \to \infty$. The eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{x}_0}$ are then given by $\alpha_1 = c$ and $\alpha_2 = 1$, such that the eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{x}_0}$ become $\lambda_1 + \alpha_1(1-\lambda_1) = \lambda_1 + c(1-\lambda_1)$ and $\lambda_2 + \alpha_2(1-\lambda_2) = 1$. (C) The same as panel B, where the uncertainty c is reduced to the transition point $c = c^*$ ($I_{\mathrm{past}} = I_{\mathrm{past}}^*$). At this point α_1 has been reduced sufficiently such that the ratio of axes lengths as set by the corresponding eigenvalues becomes equal between the two future signal distributions $p(\boldsymbol{S}_\tau|\boldsymbol{X}_0)$ and $p(\boldsymbol{S}_\tau|\boldsymbol{S}_0)$ (Eq. 2.117), i.e. the outer green and inner red ellipse have the same aspect ratio. (D) As the uncertainty is reduced beyond the transition point, the aspect ratio remains equal for both the past (left, blue) and future (right, green) distributions. In the limit of no uncertainty, $c = 0$, the distribution $p(\boldsymbol{S}_0|\boldsymbol{X}_0)$ would localize to a point (left, blue), while the future uncertainty as set by $p(\boldsymbol{S}_\tau|\boldsymbol{X}_0)$ (right, outer green) would reduce to that of the signal set by $p(\boldsymbol{S}_\tau|\boldsymbol{S}_0)$ (right, inner red). This distribution therefore sets the upper bound on the predictive information (panel A).

To make progress, let us consider the covariance matrices of the current signal given its future value $\boldsymbol{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{s}_\tau}$, and the future signal given its current value $\boldsymbol{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{s}_0}$. Again using the Schur complement formula (Eq. 2.13) these matrices can be expressed as

$$\boldsymbol{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{s}_\tau} = \mathbb{I} - \boldsymbol{\Sigma}_{\boldsymbol{s}\boldsymbol{s}_\tau}\boldsymbol{\Sigma}_{\boldsymbol{s}\boldsymbol{s}_\tau}^T, \tag{2.94}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{s}_0} = \mathbb{I} - \boldsymbol{\Sigma}_{\boldsymbol{s}\boldsymbol{s}_\tau}^T\boldsymbol{\Sigma}_{\boldsymbol{s}\boldsymbol{s}_\tau}. \tag{2.95}$$

Using these identities and the fact that $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are left eigenvectors of the matrix $\boldsymbol{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{s}_\tau}$, we can straightforwardly obtain the eigenvectors and -values of $\boldsymbol{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{s}_0}$,

$$\boldsymbol{v}_i\left(\mathbb{I} - \boldsymbol{\Sigma}_{\boldsymbol{s}\boldsymbol{s}_\tau}\boldsymbol{\Sigma}_{\boldsymbol{s}\boldsymbol{s}_\tau}^T\right) = \boldsymbol{v}_i\lambda_i, \tag{2.96}$$

$$\boldsymbol{v}_i\boldsymbol{\Sigma}_{\boldsymbol{s}\boldsymbol{s}_\tau}\left(\mathbb{I} - \boldsymbol{\Sigma}_{\boldsymbol{s}\boldsymbol{s}_\tau}^T\boldsymbol{\Sigma}_{\boldsymbol{s}\boldsymbol{s}_\tau}\right) = \boldsymbol{v}_i\boldsymbol{\Sigma}_{\boldsymbol{s}\boldsymbol{s}_\tau}\lambda_i, \tag{2.97}$$

$$\boldsymbol{v}_i\boldsymbol{\Sigma}_{\boldsymbol{s}\boldsymbol{s}_\tau}\boldsymbol{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{s}_0} = \boldsymbol{v}_i\boldsymbol{\Sigma}_{\boldsymbol{s}\boldsymbol{s}_\tau}\lambda_i, \tag{2.98}$$

where in the second line we right-multiplied with $\boldsymbol{\Sigma}_{\boldsymbol{s}\boldsymbol{s}_\tau}$ and in the last line we used Eq. 2.95. The eigenvectors of $\boldsymbol{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{s}_0}$ therefore are,

$$\boldsymbol{w}_1 = \boldsymbol{v}_1\boldsymbol{\Sigma}_{\boldsymbol{s}\boldsymbol{s}_\tau}, \tag{2.99}$$

$$\boldsymbol{w}_2 = \boldsymbol{v}_2\boldsymbol{\Sigma}_{\boldsymbol{s}\boldsymbol{s}_\tau}, \tag{2.100}$$

with corresponding eigenvalues λ_1 and λ_2.

Conveniently, the eigenvectors of $\boldsymbol{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{s}_0}$ (Eqs. 2.99 and 2.100) are also eigenvectors of the matrix $\boldsymbol{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{x}_0}$, since using Eq. 2.93 we find,

$$\boldsymbol{w}_i\boldsymbol{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{x}_0} = \boldsymbol{w}_i\left(1 - \frac{||\boldsymbol{w}_i||^2}{1+\sigma_{\xi_i}}\right), \tag{2.101}$$

where via Eq. 2.96 we have for the squared magnitude of the eigenvectors,

$$||\boldsymbol{w}_i||^2 = \boldsymbol{v}_i\boldsymbol{\Sigma}_{\boldsymbol{s}\boldsymbol{s}_\tau}\boldsymbol{\Sigma}_{\boldsymbol{s}\boldsymbol{s}_\tau}^T\boldsymbol{v}_i^T = 1 - \lambda_i. \tag{2.102}$$

We thus obtain for the eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{x}_0}$, combining Eqs. 2.101 and 2.102,

$$\lambda^{(1)}_{s_\tau|x_0} = \frac{\lambda_1+\sigma^2_{\xi_1}}{1+\sigma^2_{\xi_1}} = \lambda_1 + \alpha_1(1-\lambda_1), \tag{2.103}$$

$$\lambda^{(2)}_{s_\tau|x_0} = \frac{\lambda_2+\sigma^2_{\xi_2}}{1+\sigma^2_{\xi_2}} = \lambda_2 + \alpha_2(1-\lambda_2), \tag{2.104}$$

with α_1 and α_2, the eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{x}_0}$, given by Eqs. 2.88 and 2.89. The eigenvalues in Eqs. 2.103 and 2.104 quantify the spread of the distribution $p(\boldsymbol{S}_\tau|\boldsymbol{X}_0)$ in the directions given by the corresponding eigenvectors (Eqs. 2.99 and 2.100), as illustrated in the right panel of Fig. 2.3B. As for the uncertainty on the past signal, the maximal uncertainty on the future signal in each direction is one, and occurs when the noise in the corresponding mapping component diverges $\sigma^2_{\xi_i} \to \infty$ such that $\alpha_i = 1$ (Eqs. 2.88 and 2.89). However, where the uncertainty about the past could be reduced to zero, there is always a minimum amount of uncertainty about the future. This uncertainty corresponds to the inherent uncertainty in the evolution of the signal, and is given by the eigenvalues λ_1 and λ_2 of the signal statistics matrix $\mathbf{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{s}_0}$.

In short, the eigenvalues $\lambda^{(i)}_{s_\tau|x_0}$ (Eqs. 2.103 and 2.104) are a property of the distribution $p(\boldsymbol{S}_\tau|\boldsymbol{X}_0)$, and depend both on the signal eigenvalues λ_i of the distribution $p(\boldsymbol{S}_\tau|\boldsymbol{S}_0)$, which the system cannot optimize over, and on the eigenvalues α_i of the mapping distribution $p(\boldsymbol{S}_0|\boldsymbol{X}_0)$, which are set by the corresponding mapping noise $\sigma^2_{\xi_i}$ (Eqs. 2.88 and 2.89). Indeed, optimizing $\lambda^{(i)}_{s_\tau|x_0}$ entails optimizing for α_i over the noise magnitude $\sigma^2_{\xi_i}$.

The determinant of $\mathbf{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{x}_0}$ is the product of its eigenvalues and sets the predictive information (Eq. 2.81)

$$I_{\text{pred}}(\boldsymbol{S}_\tau;\boldsymbol{X}_0) = -\frac{1}{2}\log\left(\lambda^{(1)}_{s_\tau|x_0}\lambda^{(2)}_{s_\tau|x_0}\right) = I_1 + I_2, \tag{2.105}$$

with the predictive information in each orthogonal mapping direction,

$$I_1 = -\frac{1}{2}\log\left(\lambda^{(1)}_{s_\tau|x_0}\right) = -\frac{1}{2}\log(\lambda_1 + \alpha_1(1-\lambda_1)), \tag{2.106}$$

$$I_2 = -\frac{1}{2}\log\left(\lambda^{(2)}_{s_\tau|x_0}\right) = -\frac{1}{2}\log(\lambda_2 + \alpha_2(1-\lambda_2)), \tag{2.107}$$

where λ_1 and λ_2 are the eigenvalues of the matrices of signal statistics $\mathbf{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{s}_\tau}$ and $\mathbf{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{s}_0}$, and α_1 and α_2 are the eigenvalues of $\mathbf{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{x}_0}$ (Eqs. 2.88 and 2.89). Similar to the past information, when only one component is used in the mapping and $\sigma^2_{\xi_2} \to \infty$ such that $\alpha_2 = 1$, the predictive information depends only on the signal uncertainty and mapping noise in the direction of the first component, λ_1 and $\sigma^2_{\xi_1}$ respectively (Eq. 2.106 with α_1 of Eq. 2.88).

INVESTIGATING THE OPTIMAL STRATEGY

We now use the derivations above to obtain the optimal strategy that maximizes the predictive information given constrained past information. This derivation will elucidate why it is optimal to start out with one output component, and add an additional output component only after a given transition point.

As we have shown above, constraining the past information is equivalent to constraining the uncertainty in the current signal given the current output, as quantified by the matrix determinant $|\mathbf{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{x}_0}|$ (Eq. 2.80). In the optimum, which requires that $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are eigenvectors of both $\mathbf{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{s}_\tau}$ and $\mathbf{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{x}_0}$, this determinant is set by the pair of noise variances $(\sigma^2_{\xi_1}, \sigma^2_{\xi_2})$ (Eqs. 2.80 and 2.90). We set the constraint

$$|\mathbf{\Sigma}_{\boldsymbol{s}_0|\boldsymbol{x}_0}| = \alpha_1\alpha_2 = c, \tag{2.108}$$

which in turn sets the past information $I_{\text{past}} = -1/2\log c$ (Eq. 2.90). Via Eqs. 2.88 and 2.89, this constraint relates the noise variances $\sigma^2_{\xi_1}$ and $\sigma^2_{\xi_2}$, but it does not uniquely specify them. To this end, we need to maximize the predictive information for fixed c.

Maximizing the predictive information corresponds to minimizing the uncertainty about the future signal conditioned on the current system output. This corresponds to minimizing the total spread of the distribution $p(\boldsymbol{S}_\tau|\boldsymbol{X}_0)$, as given by $|\mathbf{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{x}_0}|$. The matrix determinant is the product of its eigenvalues $\lambda^{(1)}_{s_\tau|x_0}$ and $\lambda^{(2)}_{s_\tau|x_0}$ (Eqs. 2.103 and 2.104), which fully specify the predictive information (Eq. 2.105). As discussed below Eqs. 2.103 and 2.104, optimizing these eigenvalues entails optimizing α_1 and α_2, which govern the mapping from S_0 to X_0 (Eqs. 2.88 and 2.89). The latter are related via the constraint on the past information, such that we can substitute $\alpha_2 = c/\alpha_1$ (Eq. 2.108). The matrix determinant $|\mathbf{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{x}_0}| = \lambda^{(1)}_{s_\tau|x_0}\lambda^{(2)}_{s_\tau|x_0}$ thus becomes a function of c and α_1,

$$|\mathbf{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{x}_0}| = (\lambda_1 + \alpha_1(1-\lambda_1))(\lambda_2 + c(1-\lambda_2)/\alpha_1), \tag{2.109}$$

where again, λ_1 and λ_2 are properties of the signal that cannot be optimized (see also Eqs. 2.99 and 2.100). Taking the derivative of Eq. 2.109 with respect to α_1 and equating to zero yields the following relations at the information bound, where the predictive information is maximized for a given past information or uncertainty c:

$$\alpha_1^{\text{opt}} = \sqrt{cc^*} \quad \text{for } c \le c^*, \tag{2.110}$$

$$\alpha_2^{\text{opt}} = \sqrt{c/c^*} \quad \text{for } c \le c^*, \tag{2.111}$$

where

$$c^* = \frac{\lambda_1(1-\lambda_2)}{\lambda_2(1-\lambda_1)}. \tag{2.112}$$

Inspection of α_2^{opt} (Eq. 2.111) reveals that c^* defines the transition point: when $c = c^*$ we have $\alpha_2^{\text{opt}} = 1$ (Fig. 2.3C), which corresponds to $\sigma^2_{\xi_2} \to \infty$. Only as the uncertainty c is decreased below c^*, increasing the past information above $I^*_{\text{past}} = -1/2\log c^*$ (Fig. 2.3A, D), the noise $\sigma^2_{\xi_2}$ decreases to finite values, and the second component is added to the optimal mapping (see Eqs. 2.77-2.79). In the opposing regime, at the start of the information bound, where $c > c^*$ and $I_{\text{past}} < I^*_{\text{past}}$ (Fig. 2.3A), we have $\alpha_2^{\text{opt}} = 1$, which is its maximal value, such that we require $\alpha_1^{\text{opt}} = c$ to obey the constraint defined in Eq. 2.108 (corresponding to Fig. 2.3B). Notably, reducing the uncertainty beyond the transition point, $c < c^*$ and thus $I_{\text{past}} > I^*_{\text{past}}$, the ratio of the uncertainty in each principal direction of $p(\boldsymbol{S}_0|\boldsymbol{X}_0)$ remains fixed: $\alpha_1^{\text{opt}}/\alpha_2^{\text{opt}} = c^*$ (Eqs. 2.110 and 2.111, Fig. 2.3D).

We can gain a more intuitive understanding of the transition point c^* when we consider the eigenvalues of the matrix $\mathbf{\Sigma}_{\boldsymbol{s}_\tau|\boldsymbol{x}_0}$ in the optimum. These eigenvalues quantify the magnitude of the uncertainty in the future signal (Fig. 2.3B, right panel). Substituting Eqs. 2.110 and 2.111 in Eqs. 2.103 and 2.104 we find,

$$\lambda^{(1),\text{opt}}_{s_\tau|x_0} = \begin{cases} \lambda_1 + c(1-\lambda_1) & \text{for } c > c^*, \\ \lambda_1\left(1+\sqrt{\frac{c(1-\lambda_1)(1-\lambda_2)}{\lambda_1\lambda_2}}\right) & \text{for } c \le c^*, \end{cases} \tag{2.113}$$

$$\lambda^{(2),\text{opt}}_{s_\tau|x_0} = \begin{cases} 1 & \text{for } c > c^*, \\ \lambda_2\left(1+\sqrt{\frac{c(1-\lambda_1)(1-\lambda_2)}{\lambda_1\lambda_2}}\right) & \text{for } c \le c^*. \end{cases} \tag{2.114}$$

These optimal eigenvalues yield a few interesting observations in the distinct regimes before and after the transition point $c = c^*$.

For minimal past information $I_{\text{past}} = 0$, corresponding to maximal uncertainty $c = 1$, both eigenvalues are 1. As discussed below Eqs. 2.103 and 2.104 the minima of these eigenvalues are set by the inherent uncertainty of the signal given by λ_1 and λ_2. Therefore, the uncertainty that can be reduced in each direction is $1 - \lambda_1$ and $1 - \lambda_2$, with $\lambda_1 < \lambda_2 < 1$ (Fig. 2.3B, right panel). When $I_{\text{past}} = 0$, the predictive information is also 0, and the maximal predictive information that can be gained by each of the two components is, respectively, (following the decomposition of Eqs. 2.105, 2.106 and 2.107)

$$I_1^{\max} = -\frac{1}{2}\log\lambda_1, \tag{2.115}$$

$$I_2^{\max} = -\frac{1}{2}\log\lambda_2. \tag{2.116}$$

The most information can thus be gained by using the first mapping component, $I_1^{\max} > I_2^{\max}$. Therefore, as the past information is increased, and correspondingly the uncertainty c is reduced, the optimal system initially uses only the first component (Fig. 2.3A and B). In this component the available coding space is the largest.

When the past information is increased to the transition point $I_{\text{past}} = I^*_{\text{past}} = -1/2\log c^*$ (i.e. $c = c^*$), an interesting observation can be made: at this point the ratio of the optimal eigenvalues of $p(\boldsymbol{S}_\tau|\boldsymbol{X}_0)$ (Eqs. 2.113 and 2.114) starts to equal that of the inherent signal uncertainty $p(\boldsymbol{S}_\tau|\boldsymbol{S}_0)$, i.e.,

$$\lambda^{(1),\text{opt}}_{S_\tau|X_0} / \lambda^{(2),\text{opt}}_{S_\tau|X_0} = \lambda_1/\lambda_2 \quad \text{for } c \le c^*, \tag{2.117}$$

see also Fig. 2.3C. What this means becomes more apparent when we consider the predictive information at the transition point. Using the definition of the predictive information (Eq. 2.105) and the optimal eigenvalues (Eqs. 2.113 and 2.114) we obtain

$$I^* = -\frac{1}{2}\left(\log\lambda_1 - \log\lambda_2\right). \tag{2.118}$$

Interestingly, at this point, the remaining predictive information accessible via the first component is equal to the total predictive information associated with the second component,

$$I_1^{\max} - I^* = I_2^{\max} = -\frac{1}{2}\log\lambda_2, \tag{2.119}$$

see Fig. 2.3A. From this point onward the system should use both mapping components equally, because there is equal coding space available in either (Fig. 2.3C). What is more, for each value of the past information beyond the transition point, $I_{\text{past}} > I^*_{\text{past}}$ corresponding to $c < c^*$ (Fig. 2.3D), the ratio of eigenvalues remains equal (Eq. 2.117) and the predictive information contributed by each component relative to their value at the transition point is the same. Indeed, using the predictive information at the transition point (Eq. 2.118), and the decomposed predictive information Eqs. 2.106 and 2.107 with the optimal eigenvalues Eqs. 2.113 and 2.114, we obtain

$$I_1 - I^* = I_2 = -\frac{1}{2}\log\left(\lambda_2\left[1 + \sqrt{\frac{c(1-\lambda_1)(1-\lambda_2)}{\lambda_1\lambda_2}}\right]\right), \quad \text{for } c \le c^*. \tag{2.120}$$

This is further illustrated in Fig. 2.3A.

In summary, to maximize its predictive information under highly constrained past information, the optimal system first uses only the output component with the largest available coding space, which thus has the largest potential for uncertainty reduction (Fig. 2.3B). Using this component increases the predictive information the fastest as function of the past information. This is the optimal strategy until the point that the available encoding space, i.e. the remaining uncertainty, becomes equal in both coding directions (Fig. 2.3C). From this point onwards, the optimal way to increase the predictive information as a function of the past information, is by increasing the predictive information in both directions by the same amount, such that the remaining free coding space remains equal in both directions (Fig. 2.3A, D).

The analysis above is reminiscent of a classical and perhaps less abstract problem in physical chemistry, asking the question how one should distribute non-overlapping particles over two boxes of unequal volume, such that the total free energy of the system remains minimal. The answer is quite intuitive: one should add particles to the largest box until its remaining volume equals that of the smaller box. After this point the particles must be distributed equally over both boxes, such that the free volume in the two boxes remains equal. This procedure keeps the chemical potentials for the particles in the two boxes equal, which minimizes the total free energy of the system.

2.7. Discussion

In this chapter we have studied how systems should optimally extract information from the past to maximize the information they obtain about the future. We found that, for signals that are Markovian in their concentration, the optimal responder simply copies the most recent input into its output. This reflects the Markovian nature of the signal: its past does not contain any additional information about its future that is not already contained in the current signal. When we add a noise term to the input signal that contains no information about its future state, the signal becomes non-Markovian and the optimal system must use signal values further in the past to average out the noise.

For a slightly more complex signal that is defined by two features, both its current concentration value and derivative, we find that the optimal prediction strategy generally exploits both properties. For example, to predict the future derivative, the optimal system bases its current output on a linear combination of the concentration value and derivative, proportional to their correlation with the future derivative. The optimal strategy is again instantaneous, because the signal is Markovian in the joint signal value and derivative. When a system must predict both the concentration value and the derivative of such a signal, it should, for sufficient past information, start to use multiple output components. This transition point occurs when the remaining predictive information associated with each optimal encoding direction is equal.

In the chapters that follow we will study how the optimal strategies derived here translate to biology. Can single-celled organisms implement them using biochemical signaling networks, and what limitations do they face? It is evident from this chapter that past information is a resource that is required to obtain predictive information, but how does it relate to resources that have a direct physical cost, such as proteins, energy, and time?

APPENDIX

2.A. CAUSAL WIENER FILTER

Here we present a brief derivation of the causal Wiener filter for optimal prediction, combining elements of the derivations presented in [7, 44, 46]. We consider a jointly Gaussian degraded past signal $s(t)$ and future signal of interest $\ell_\tau = \ell(t+\tau)$; both are shifted to have zero mean. The objective is to find the causal integration kernel $k(t)$ that minimizes the mean squared prediction error between the true future signal ℓ_τ, and the output $x(t) = \hat{\ell}_\tau$ of a system that estimates the future signal:

$$\min_{k(t)} \langle \epsilon^2 \rangle = \langle (\ell_\tau - x(t))^2 \rangle, \tag{2.121}$$

where

$$x(t) = \int_{-\infty}^{t} k(t-t')s(t')dt' = \int_0^\infty k(t'')s(t-t'')dt''. \tag{2.122}$$

Expanding the mean squared prediction error and substitution of Eq. 2.122 for $x(t)$ yields,

$$\begin{aligned}\langle \epsilon^2 \rangle &= \sigma_\ell^2 - 2\langle \ell(t+\tau)x(t)\rangle + \sigma_x^2, &(2.123)\\ &= \sigma_\ell^2 - 2\int_0^\infty k(t')\langle \ell(t+\tau)s(t-t')\rangle dt' + \int_0^\infty\int_0^\infty k(t')k(t'')\langle s(t-t')s(t-t'')\rangle dt'dt''.\end{aligned}$$

For simplicity of notation let us define $C_{s\ell}(t'+\tau) = \langle \ell(t+\tau)s(t-t')\rangle$ and $C_s(t''-t') = \langle s(t-t')s(t-t'')\rangle$. Taking the functional derivative with respect to the integration kernel we obtain,

$$\frac{\delta\langle \epsilon^2\rangle}{\delta k} = -2C_{s\ell}(t'+\tau) + 2\int_0^\infty k(t'')C_s(t''-t')dt''. \tag{2.124}$$

For the optimal kernel $k_*(t)$ the following relation must hold,

$$\int_0^\infty k_*(t')C_s(t'-t)dt' = C_{s\ell}(t+\tau). \tag{2.125}$$

To make progress we need to transition from the time domain to the frequency domain using the Fourier transform. In this work our convention for the Fourier transform and its inverse is as follows,

$$K(\omega) = \mathcal{F}\{k(t)\} = \int_{-\infty}^{\infty} k(t)e^{-i\omega t}dt, \tag{2.126}$$

$$k(t) = \mathcal{F}^{-1}\{K(\omega)\} = \frac{1}{2\pi}\int_{-\infty}^{\infty} K(\omega)e^{i\omega t}d\omega. \tag{2.127}$$

As is apparent from these definitions, the Fourier transform is two-sided: it requires integration over all positive and negative times and frequencies. This poses a challenge in the derivation of the causal Wiener filter, which requires the kernel $k(t)$ to be causal, i.e.

to strictly act on positives times t (which translates to positive time delays $t-t'$, see Eq. 2.122). To prudently track the causality of the derived filter we define the causal part of a function in the frequency domain as,

$$[K(\omega)]_+ = \mathcal{F}\{\theta(t)k(t)\} = \int_{-\infty}^{\infty} \theta(t)k(t)e^{-i\omega t}dt, \tag{2.128}$$

where $\theta(t)$ is the unit step function. Moreover, we will require the Wiener-Hopf factorization of a power spectrum $S(\omega)$ (power spectral density), which is defined as

$$m_+(\omega)m_-(\omega) = S(\omega), \tag{2.129}$$

where $m_+(\omega) = m_-^*(\omega) = m_-(-\omega)$. Both are real functions in the time domain, but the factorization is such that $m_+(\omega)$ and $1/m_+(\omega)$ are fully causal, while $m_-(\omega)$ and $1/m_-(\omega)$ are anti-causal, i.e. they have no causal part. For simple functions in the frequency domain the Wiener-Hopf factorization can often be identified relatively easily. Consider for example an exponential function $\exp(-\mu t)$ with Fourier transform $(\mu^2+\omega^2)^{-1}$. The Wiener-Hopf factorization of this function is $m_+(\omega) = (\mu+i\omega)^{-1}$ and $m_-(\omega) = (\mu-i\omega)^{-1}$.

We can now continue from Eq. 2.125. Let us first consider that we can express the one-sided integral over t' as follows,

$$\begin{aligned}\int_0^{\infty} k_*(t')C_s(t'-t)dt' &= \int_{-\infty}^{\infty} \theta(t)k_*(t')C_s(t'-t)dt',\\ &= \int_{-\infty}^{\infty} k_*(t')C_s(t'-t)dt' - a(t),\end{aligned} \tag{2.130}$$

where $a(t)$ is a function that eliminates the anti-causal part of the two-sided integral. Therefore, $a(t)$ itself is anti-causal, i.e. it has some defined but unknown value for $t<0$ and is zero otherwise. Substitution of Eq. 2.130 in Eq. 2.125 yields,

$$\int_{-\infty}^{\infty} k_*(t')C_s(t'-t)dt' = C_{s\ell}(t+\tau) + a(t). \tag{2.131}$$

Taking the Fourier transform of both sides enables us to exploit the convolution theorem,

$$K_*(\omega)S_s(\omega) = S_{s\ell}(\omega)e^{i\omega\tau} + A(\omega), \tag{2.132}$$

where we note that by the Wiener-Khinchin theorem the Fourier transform of a correlation function $C_x(t)$ is its power spectrum $S_x(\omega)$, and we exploited that we can express the time-shift caused by the prediction interval τ as $\mathcal{F}\{C_{s\ell}(t+\tau)\} = S_{s\ell}(\omega)\exp(i\omega\tau)$. We use the Wiener-Hopf factorization of $S_s(\omega) = m_+(\omega)m_-(\omega)$ to ensure the left-hand side is causal,

$$K_*(\omega)m_+(\omega) = \frac{S_{s\ell}(\omega)}{m_-(\omega)}e^{i\omega\tau} + \frac{A(\omega)}{m_-(\omega)}. \tag{2.133}$$

Finally, taking the causal part of both sides as in Eq. 2.128 and rearranging yields the optimal causal filter that minimizes the mean squared prediction error,

$$K_*(\omega) = \frac{1}{m_+(\omega)}\left[\frac{S_{s\ell}(\omega)}{m_-(\omega)}e^{i\omega\tau}\right]_+. \tag{2.134}$$

Here, the term $A(\omega)/m_-(\omega)$ vanished because it is anti-causal, i.e. it has no causal part.

2.B. Linear mapping to normalize input signal

Here we demonstrate how to normalize the input signal such that its covariance matrix becomes the identity matrix. We then continue to show how each relevant matrix of signal statistics can be mapped between the different bases, and finally how the rescaled optimal mapping matrix can be mapped back onto the original signal basis.

Consider a stationary Gaussian multivariate signal $\boldsymbol{S}$ that is defined by its covariance matrix

$$\boldsymbol{\Sigma_s} = \mathbb{E}\left[\boldsymbol{S}\boldsymbol{S}^T\right]. \tag{2.135}$$

This is a full-rank, real, symmetric matrix. Because the mutual information between Gaussian random variables is independent of their mean values (Eq. 2.7), these can be set to zero without loss of generality.

Our aim is to define a rescaled signal $\hat{\boldsymbol{S}}$ such that its covariance matrix is the identity matrix,

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{s}}} = \mathbb{E}\left[\hat{\boldsymbol{S}}\hat{\boldsymbol{S}}^T\right] = \mathbb{1}_{\mathrm{dim}(\boldsymbol{s})}. \tag{2.136}$$

To arrive at the appropriate scaling, let us consider that the original covariance matrix $\boldsymbol{\Sigma_s}$ can be diagonalized by the matrix of its normalized eigenvectors $\boldsymbol{P}$:

$$\boldsymbol{\Sigma_s} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^T, \tag{2.137}$$

where by orthogonality $\boldsymbol{P}^T\boldsymbol{P} = \mathbb{1}_{\mathrm{dim}(\boldsymbol{s})}$, and $\boldsymbol{D} = \mathrm{diag}(\lambda_s^{(1)}, \lambda_s^{(2)}, \ldots, \lambda_s^{\mathrm{dim}(\boldsymbol{s})})$ is a diagonal matrix with entries given by the eigenvalues of $\boldsymbol{\Sigma_s}$.

Equation 2.137 suggests a definition of $\hat{\boldsymbol{S}}$ via the following rescaling of the original signal,

$$\hat{\boldsymbol{S}} = \boldsymbol{M}\boldsymbol{S}; \quad \boldsymbol{M} = \boldsymbol{D}^{-1/2}\boldsymbol{P}^T. \tag{2.138}$$

Indeed, substituting this definition of $\hat{\boldsymbol{S}}$ in Eq. 2.136 yields the desired result,

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{s}}} = \boldsymbol{M}\boldsymbol{\Sigma_s}\boldsymbol{M}^T \tag{2.139}$$

$$= \boldsymbol{D}^{-1/2}\boldsymbol{P}^T\boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^T\boldsymbol{P}\boldsymbol{D}^{-1/2}, \tag{2.140}$$

$$= \mathbb{1}_{\mathrm{dim}(\boldsymbol{s})}. \tag{2.141}$$

Since the signal is stationary, this scaling (Eq. 2.138) holds for the signal at any point in time. We thus also have $\boldsymbol{\Sigma}_{\hat{\boldsymbol{s}}_\tau} = \boldsymbol{M}\boldsymbol{\Sigma}_{\boldsymbol{s}_\tau}\boldsymbol{M}^T = \mathbb{1}$.

For the cross-correlation matrix from past to future we simply obtain,

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{s}}\hat{\boldsymbol{s}}_\tau} = \mathbb{E}\left[\hat{\boldsymbol{S}}(0)\hat{\boldsymbol{S}}(\tau)^T\right], \tag{2.142}$$

$$= \boldsymbol{M}\Sigma_{\boldsymbol{s}\boldsymbol{s}_\tau}\boldsymbol{M}^T. \tag{2.143}$$

And the cross-correlation matrix from future to past is given by the transpose $\boldsymbol{\Sigma}_{\hat{\boldsymbol{s}}_\tau\hat{\boldsymbol{s}}} = \boldsymbol{\Sigma}^T_{\hat{\boldsymbol{s}}\hat{\boldsymbol{s}}_\tau}$.

The final matrices of signal statistics to consider are the current signal covariance conditioned on the future signal, and vice versa, the future signal covariance conditioned on the current signal. We have, exploiting the Schur complement formula (Eq. 2.13),

$$\begin{aligned}\boldsymbol{\Sigma}_{\hat{\boldsymbol{s}}|\hat{\boldsymbol{s}}_\tau} &= \boldsymbol{\Sigma}_{\hat{\boldsymbol{s}}} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{s}}\hat{\boldsymbol{s}}_\tau}\boldsymbol{\Sigma}^{-1}_{\hat{\boldsymbol{s}}_\tau}\boldsymbol{\Sigma}^T_{\hat{\boldsymbol{s}}\hat{\boldsymbol{s}}_\tau}, &(2.144)\\ &= \boldsymbol{M}\boldsymbol{\Sigma}_{\boldsymbol{s}}\boldsymbol{M}^T - \boldsymbol{M}\boldsymbol{\Sigma}_{\boldsymbol{s}\boldsymbol{s}_\tau}\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{\Sigma}^T_{\boldsymbol{s}\boldsymbol{s}_\tau}\boldsymbol{M}^T, &(2.145)\end{aligned}$$

where we used that $\boldsymbol{\Sigma}^{-1}_{\hat{\boldsymbol{s}}_\tau} = \mathbb{I}$. To make progress, let us consider the inverse of the original signal covariance matrix, starting from Eq. 2.137,

$$\begin{aligned}\boldsymbol{\Sigma}^{-1}_{\boldsymbol{s}} &= (\boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^T)^{-1}, &(2.146)\\ &= (\boldsymbol{P}^T)^{-1}\boldsymbol{D}^{-1}(\boldsymbol{P})^{-1}, &(2.147)\\ &= \boldsymbol{P}\boldsymbol{D}^{-1}\boldsymbol{P}^T = \boldsymbol{M}^T\boldsymbol{M}. &(2.148)\end{aligned}$$

Due to stationarity we then also have $\boldsymbol{\Sigma}^{-1}_{\boldsymbol{s}_\tau} = \boldsymbol{\Sigma}^{-1}_{\boldsymbol{s}} = \boldsymbol{M}^T\boldsymbol{M}$. Using this identity in Eq. 2.145 we obtain

$$\begin{aligned}\boldsymbol{\Sigma}_{\hat{\boldsymbol{s}}|\hat{\boldsymbol{s}}_\tau} &= \boldsymbol{M}\boldsymbol{\Sigma}_{\boldsymbol{s}}\boldsymbol{M}^T - \boldsymbol{M}\boldsymbol{\Sigma}_{\boldsymbol{s}\boldsymbol{s}_\tau}\boldsymbol{\Sigma}^{-1}_{\boldsymbol{s}_\tau}\boldsymbol{\Sigma}^T_{\boldsymbol{s}\boldsymbol{s}_\tau}\boldsymbol{M}^T, &(2.149)\\ &= \boldsymbol{M}\Sigma_{\boldsymbol{s}|\boldsymbol{s}_\tau}\boldsymbol{M}^T. &(2.150)\end{aligned}$$

By analogy we can also see that $\boldsymbol{\Sigma}_{\hat{\boldsymbol{s}}_\tau|\hat{\boldsymbol{s}}} = \boldsymbol{M}\Sigma_{\boldsymbol{s}_\tau|\boldsymbol{s}}\boldsymbol{M}^T$.

Finally, let us consider the mapping of the rescaled signal onto a rescaled system output $\hat{\boldsymbol{X}}$. For the original signal we have,

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S} + \boldsymbol{\xi}, \tag{2.151}$$

while for the rescaled signal we define,

$$\hat{\boldsymbol{X}} = \hat{\boldsymbol{A}}\hat{\boldsymbol{S}} + \hat{\boldsymbol{\xi}}. \tag{2.152}$$

By comparison between Eqs. 2.151 and 2.152 and using the definition of $\hat{\boldsymbol{S}}$ it is apparent that by defining $\hat{\boldsymbol{\xi}} = \boldsymbol{\xi}$ and $\hat{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{M}^{-1}$ we find,

$$\hat{\boldsymbol{X}} = \boldsymbol{A}\boldsymbol{M}^{-1}\boldsymbol{M}\boldsymbol{S} + \boldsymbol{\xi} = \boldsymbol{X}. \tag{2.153}$$

Since the mapping from $\boldsymbol{S}$ to $\hat{\boldsymbol{S}}$ is deterministic, and with the definitions above we have $\hat{\boldsymbol{X}} = \boldsymbol{X}$, both the past and predictive information are equivalent between the original and the rescaled signal basis, i.e.

$$\begin{aligned}I(\hat{\boldsymbol{S}}_0;\hat{\boldsymbol{X}}_0) &= I(\hat{\boldsymbol{S}}_0;\boldsymbol{X}_0) = I(\boldsymbol{S}_0;\boldsymbol{X}_0), &(2.154)\\ I(\hat{\boldsymbol{S}}_\tau;\hat{\boldsymbol{X}}_0) &= I(\hat{\boldsymbol{S}}_\tau;\boldsymbol{X}_0) = I(\boldsymbol{S}_\tau;\boldsymbol{X}_0). &(2.155)\end{aligned}$$

To obtain the mapping kernels in the original signal basis from the mapping kernels in the rescaled basis one must thus right-multiply with the matrix $\boldsymbol{M}$ of Eq. 2.138:

$$\boldsymbol{A} = \hat{\boldsymbol{A}}\boldsymbol{M}. \tag{2.156}$$

3

Not all bits are equally predictive, nor costly

Living cells can leverage correlations in environmental fluctuations to predict the future environment and mount a response ahead of time. To this end, cells need to encode the past signal into the output of the intracellular network from which the future input is predicted. Yet, storing information is costly while not all features of the past signal are equally informative on the future input signal. Here, we show that push-pull networks can reach the information bound for Markovian signals. However, the bits of past information that are most informative about the future signal are also prohibitively costly. As a result, the optimal system that maximizes the predictive information for a given resource cost is not at the information bound.

The contents of this chapter have been published in Proceedings of the National Academy of Sciences **120**, e2303078120 (2023) [38].

One of the most remarkable abilities of life is the capacity of organisms to predict the future. Although commonly associated with higher organisms, even single cells can anticipate changes in their environment [5, 6]. From an evolutionary perspective, this makes sense: cells must adapt to environmental changes, but mounting a response takes time. Therefore, when environmental fluctuations follow a temporal pattern, it is advantageous for cells to predict these changes, and mount a response ahead of time.

In Chapter 2, we have explored the fundamental class of Markovian signals. These signals are notoriously difficult to predict since their future state depends solely on their present state, with no additional information about the future encoded in the past. As one might therefore expect, we find that for such signals, the optimal system that maximizes the predictive information under constrained past information, copies only the current signal into the output from which it predicts the future. This leads us to hypothesize that a ubiquitous biochemical signaling motif, the push-pull network, could reach the information bound, as it effectively copies the current signal value into its output [13, 15, 47]. However, it remains unclear whether this network can reach the bound when biophysical resource constraints such as limited protein copies and energy are taken into account. And even if it could: is being at the information bound truly optimal? How close do networks designed to maximize the predictive information under a biophysical resource constraint come to this bound?

In what follows, we first revisit key concepts from linear signaling theory and the Gaussian information bottleneck. Next, we define the input signal considered in this chapter and summarize several key results from Chapter 2. We then derive general expressions for the past and predictive information in linear signaling systems before introducing a model of the push-pull network. Although this network is able to reach the information bound, our results show that the optimal push-pull network that maximizes the predictive information under a biophysical resource constraint set by protein copies and energy is not at the bound. In the sections that follow, we delve into why this is the case. At the core of our findings lies a perhaps simple but fundamental insight: not all bits of information are equally predictive, nor equally costly. Ultimately, it is the tradeoff between those bits of information that are most predictive, and those bits that are cheapest, that determines the optimal design of the push-pull network.

3.1. LINEAR NETWORKS AND THE INFORMATION BOTTLENECK METHOD

We study cellular signaling systems within their linear response regime [13, 15, 20, 34, 48]. Linear systems not only allow for analytical results, but also describe information transmission often remarkably well [31–34], also see Chapter 6. The output of these systems can be written as

$$x(t) = \int_{-\infty}^{t} dt' k(t-t')\ell(t') + \eta_x(t), \tag{3.1}$$

where $k(t)$ is the linear response function, $\ell(t)$ the input signal, and $\eta_x(t)$ describes the noise in the output. Throughout this dissertation we consider stationary signals with different temporal correlations, obeying Gaussian statistics.

In this chapter we assume that the cell needs to predict the value of the input $\ell_\tau \equiv \ell(t+\tau)$ at a time τ into the future. We expect the forecast interval τ to be set by the time to mount a response. For accurate prediction it cannot be much longer than the correlation time of the input signal.

Any prediction about the future state of the environment must be based on information obtained from its past (Fig. 1.1A). In particular, the cell needs to predict ℓ_τ from the current output $x_0 \equiv x(t)$, which itself depends on the input signal trajectory in the past, $\boldsymbol{L}_p \equiv \{\ell(t')\}_{t'<t}$. The (qualitative) shape of the integration kernel $k(t)$, e.g. exponential, adaptive or oscillatory, is determined by the topology of the signaling network [7]. The kernel shape describes how the past signal is mapped onto the current output, and hence which characteristics of the past signal the cell uses to predict the future signal. To maximize the prediction accuracy, the cell should extract those features that are most informative about the future signal. These depend on the signal statistics.

Deriving the upper bound on the predictive information as set by the past information is an optimization problem, which can be solved using the information bottleneck method (IBM) [9]. In Chapter 2 we discussed this approach in detail; here we shortly outline its main components. The IBM entails the maximization of an objective function $\mathcal{L}$:

$$\max_{P(x_0|\boldsymbol{L}_p)} \left[\mathcal{L} \equiv I(x_0;\ell_\tau) - \gamma I(x_0;\boldsymbol{L}_p)\right]. \tag{3.2}$$

Here, $I_{\text{pred}} \equiv I(x_0;\ell_\tau)$ is the predictive information, which is the mutual information between the system's current output x_0 and the future ligand concentration ℓ_τ. It quantifies the degree to which knowledge of the output x_0 reduces the uncertainty about the future input ℓ_τ, and, within the Gaussian framework, it is related to the mean-squared prediction error as derived, e.g., using Wiener filtering (see also Appendix 2.A) [7]. The past information $I_{\text{past}} \equiv I(x_0;\boldsymbol{L}_p)$ is the mutual information between x_0 and the trajectory of past ligand concentrations $\boldsymbol{L}_p$. The Lagrange multiplier γ sets the relative cost of storing past over obtaining predictive information. Given a value of γ, the objective function in Eq. 3.2 is maximized by optimizing the conditional probability distribution of the output given the past input trajectory, $P(x_0|\boldsymbol{L}_p)$. For the linear systems considered here, this corresponds to optimizing the mapping of the past input signal onto the current output via the integration kernel $k(t)$; in fact, this linear-response strategy is optimal for signals that obey Gaussian statistics [16]. In Chapter 2 we used the Gaussian IBM to derive, for different input statistics, the optimal kernel $k^{\text{opt}}(t)$ and the *information bound*, defined to be the maximum predictive information as set by the past information [16].

3.2. OPTIMAL PREDICTION OF MARKOVIAN SIGNALS: BIOCHEMICAL COPYING

Arguably the most elementary signal type is a Markovian signal $\ell(t)$ with correlation time τ_ℓ. While the signal is Markovian, its fluctuations remain correlated on the finite timescale set by τ_ℓ, which means that its future value ℓ_τ can be predicted with some accuracy. The deviations $\delta\ell(t) = \ell(t) - \bar{\ell}$ from its mean $\bar{\ell}$ follow an Ornstein-Uhlenbeck process:

$$\delta\dot{\ell} = -\delta\ell(t)/\tau_\ell + \eta_\ell(t), \tag{3.3}$$

where $\eta_\ell(t)$ is Gaussian white noise, $\langle \eta_\ell(t)\eta_\ell(t')\rangle = 2\sigma_\ell^2/\tau_\ell\delta(t-t')$, with σ_ℓ^2 the amplitude of the signal fluctuations. This input signal obeys Gaussian statistics, characterized by $\langle\delta\ell(0)\delta\ell(t)\rangle = \sigma_\ell^2\exp(-t/\tau_\ell)$. Employing the Gaussian IBM framework [16], we find that the optimal integration kernel is given by (see Section 2.3)

$$k^{\text{opt}}(t-t') = a\delta(t-t'). \tag{3.4}$$

This kernel corresponds to a signaling system that copies the current input into the output. This is intuitive, since for a Markovian signal there is no additional information in the past signal that is not already contained in the present one. The prefactor a determines the static gain $\partial\bar{x}/\partial\bar{\ell}$, which together with the noise strength $\sigma_{\eta_x}^2$ (Eq. 3.1) and the signal amplitude σ_ℓ^2 set the magnitude of the past and predictive information, I_{past} and I_{pred}, respectively (Section 2.2).

Fig. 3.2-I shows the maximum predictive information as set by the past information (black curve). This information bound applies to any system that needs to predict a Markovian signal obeying Gaussian statistics. How close can biochemical systems come to this bound?

3.3. PAST AND PREDICTIVE INFORMATION FOR LINEAR SIGNALING NETWORKS

In order to address whether biochemical networks can approach the information bound, we here describe how we obtain the past and predictive information for any linear signaling network. We then use the resulting general expressions to compute the past and predictive information for the push-pull network and compare it to the information bound for a Markovian input signal.

For any linear network the output can be written as in Eq. 3.1, where the mapping kernel $k(t)$ is a property of the network and describes how the input signal is mapped onto the output. The noise term $\eta_x(t)$ is a sum of convolutions over all white noise processes in the network and corresponding network mapping functions (see Appendix 3.C, Eq. 3.63). The variance in the output of a linear network can be split up in a part caused by the signal and a part cause by the noise, we have

$$\begin{aligned}\sigma_x^2 &= \int_{-\infty}^{t} dt' \int_{-\infty}^{t} dt'' k(t-t')k(t-t'')\langle\delta\ell(t')\delta\ell(t'')\rangle + \sigma_{\eta_x}^2,\\ &= \sigma_{x|\eta}^2 + \sigma_{x|L}^2,\end{aligned} \tag{3.5}$$

where $\sigma_{x|\eta}^2$ is the signal variance, i.e. the variance due to the signal variations and hence with the noise terms fixed, and $\sigma_{x|L}^2$ is the noise variance, i.e. with the complete history of the signal fixed.

Using the decomposition of Eq. 3.5 and the Gaussian mutual information (Eq. 2.7), we find for the past information, which is the mutual information between the current output and the complete signal history,

$$
\begin{aligned}
I_{\text{past}}(x_0;\boldsymbol{L}_p) &= \frac{1}{2}\log\left(\frac{\sigma_x^2}{\sigma_{x|L}^2}\right),\\
&= \frac{1}{2}\log\left(\frac{\sigma_{x|\eta}^2+\sigma_{x|L}^2}{\sigma_{x|L}^2}\right) = \frac{1}{2}\log(1+\text{SNR}_{\text{past}}),
\end{aligned}
\tag{3.6}
$$

where the signal-to-noise ratio (SNR) for the past information is defined as the output variance caused by any past signal fluctuations over the network noise: $\text{SNR}_{\text{past}} = \sigma_{x|\eta}^2/\sigma_{x|L}^2$.

We can similarly define the predictive information between current output and future ligand concentration,

$$
I_{\text{pred}}(x_0;\ell_\tau) = \frac{1}{2}\log\left(\frac{\sigma_x^2}{\sigma_{x|\ell_\tau}^2}\right),
\tag{3.7}
$$

where $\sigma_{x|\ell_\tau}^2$ is the variance in the output x given the future value of the input ℓ_τ. Using the Schur complement formula, Eq. 2.13, we can write the variance in the output as

$$
\begin{aligned}
\sigma_x^2 &= \sigma_{x|\ell_\tau}^2 + \langle\delta x(0)\delta\ell(\tau)\rangle^2/\sigma_\ell^2, &(3.8)\\
&= \sigma_{x|\ell_\tau}^2 + \tilde{g}^2\sigma_\ell^2, &(3.9)
\end{aligned}
$$

which defines the dynamic gain [37]

$$
\tilde{g} \equiv \langle\delta x(0)\delta\ell(\tau)\rangle/\sigma_\ell^2.
\tag{3.10}
$$

Combining Eqs. 3.7-3.10 then yields for the predictive information:

$$
I_{\text{pred}}(x_0;\ell_\tau) = \frac{1}{2}\log\left(\frac{\sigma_{x|\ell_\tau}^2+\tilde{g}^2\sigma_\ell^2}{\sigma_{x|\ell_\tau}^2}\right) = \frac{1}{2}\log(1+\text{SNR}_{\text{pred}}),
\tag{3.11}
$$

where SNR_{pred} is the ratio of the signal,

$$
\text{SIGNAL}_{\text{pred}} = \tilde{g}^2\sigma_\ell^2,
\tag{3.12}
$$

over the noise,

$$
\text{NOISE}_{\text{pred}} = \sigma_{x|\ell_\tau}^2.
\tag{3.13}
$$

Within the Gaussian framework, the inverse of the SNR_{pred} also yields the relative error in estimating the future concentration [15]:

$$
\text{SNR}_{\text{pred}}^{-1} = \left(\delta\hat{\ell}_\tau\right)^2/\sigma_\ell^2.
\tag{3.14}
$$

Equations 3.7-3.12 show that this relative error is also related to the Pearson correlation coefficient $\rho_{x\ell}(\tau)$ between the current output and the future ligand concentration:

$$
\begin{aligned}
\rho_{x\ell}(\tau) &\equiv \langle\delta x(0)\delta\ell(\tau)\rangle/(\sigma_x\sigma_\ell), &(3.15)\\
\text{SNR}_{\text{pred}}^{-1} &= \rho_{x\ell}^{-2}(\tau)-1, &(3.16)
\end{aligned}
$$

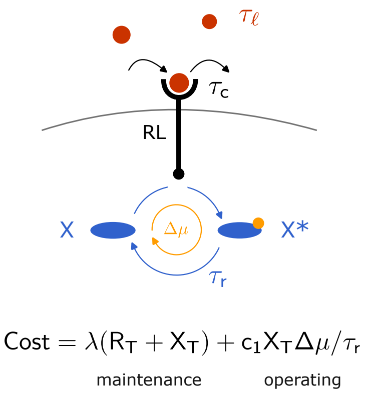


Figure 3.1: A biochemical copying device: the push-pull network. The optimal topology of the network for predicting the future signal depends on the temporal statistics of the input signal. Push-pull networks, consisting of chemical modification cycles or GTPase cycles, can optimally predict the future value of Markovian signals (Eq. 3.3), with correlation time τ_ℓ. The push-pull network we study consists of a receptor that drives a downstream phosphorylation cycle. The ligand binds the receptor with a correlation time τ_c. The push-pull network, driven by ATP hydrolysis with free energy $\Delta\mu$, integrates the receptor with an integration time τ_r. The total resource cost consists of a maintenance cost of receptor and readout synthesis at the growth rate λ, and an operating cost of driving the cycle.

such that the predictive information is also given by

$$I_{\text{pred}}(x_0;\ell_\tau) = \frac{1}{2}\log\left(1+\text{SNR}_{\text{pred}}\right) = -\frac{1}{2}\log\left(1-\rho^2_{x\ell}(\tau)\right). \tag{3.17}$$

To compute the past information via Eq. 3.6 we thus need to compute the output variance σ_x^2 in terms of its two constituent parts $\sigma^2_{x|\eta}$ and $\sigma^2_{x|L}$. To compute the predictive information with the future ligand concentration via Eq. 3.11 or Eq. 3.17, we further need to compute the covariance $\langle \delta x(0)\delta\ell(\tau)\rangle$, which gives the dynamic gain $\tilde{g}$ (Eq. 3.10).

3.4. THE PUSH-PULL NETWORK

Although the upper bound on the accuracy of prediction is determined by the signal statistics, how close cells can come to this bound depends on the topology of the cellular signaling system, and the resources devoted to building and operating it. A network motif that may reach the information bound for Markovian signals is the push-pull network (Fig. 3.1), because it is at heart a copying device: it samples the input by copying the state of the input into the activation state of the output [13, 15, 47].

We consider a push-pull network that consists of a phosphorylation-dephosphorylation cycle downstream of a receptor (Fig. 3.1). When bound to ligand, the receptor itself or its associated kinase, such as CheA in the *Escherichia coli* chemotaxis network, catalyzes the phosphorylation of a readout protein X, like CheY. Active readout molecules X^* can decay spontaneously or be deactivated by an enzyme (phosphatase), such as CheZ in *E. coli*. This cycle is driven by the turnover of fuel such as ATP. We recognize that inside the living cell, the chemical driving is typically large: for example, the free energy of ATP hydrolysis is about $20k_\mathrm{B}T$, which means that the system essentially operates in the irreversible regime [13, 15]. This system then consists of the following reactions:

$$\mathrm{R}+\mathrm{L} \underset{\mathrm{k}_-}{\overset{\mathrm{k}_+}{\rightleftharpoons}} \mathrm{RL} \tag{3.18}$$

$$\mathrm{RL}+\mathrm{X} \xrightarrow{\mathrm{k}_\mathrm{f}} \mathrm{RL}+\mathrm{X}^* \tag{3.19}$$

$$\mathrm{X}^* \xrightarrow{\mathrm{k}_\mathrm{r}} \mathrm{X} \tag{3.20}$$

Both the total number of receptors $R_\mathrm{T} = R + RL$ and read-out molecules $X_\mathrm{T} = X + X^*$ are conserved moieties. The chemical Langevin equations of this system are (also see Appendix 3.A):

$$\dot{RL} = [R_\mathrm{T} - RL(t)]\ell(t)k_+ - RL(t)k_- + B_c(RL,\ell)\xi_c(t), \tag{3.21}$$

$$\dot{x}^* = [X_\mathrm{T} - x^*(t)]RL(t)k_\mathrm{f} - x^*(t)k_\mathrm{r} + B_x(RL,x^*)\xi_x(t), \tag{3.22}$$

where RL is the number of bound receptors, x^* the number of phosphorylated readout molecules, and ξ_i denote independent Gaussian white noise with unit variance, $\langle\xi_i(t)\xi_j(t')\rangle = \delta_{ij}\delta(t-t')$. The noise strengths are (Appendix 3.A)

$$B_c(RL,\ell) = \sqrt{(R_\mathrm{T} - RL(t))\ell(t)k_+ + RL(t)k_-}, \tag{3.23}$$

$$B_x(RL,x^*) = \sqrt{(X_\mathrm{T} - x^*(t))RL(t)k_\mathrm{f} + x^*(t)k_\mathrm{r}}. \tag{3.24}$$

The steady-state fraction of ligand-bound receptors is $p \equiv \overline{RL}/R_\mathrm{T} = \bar{\ell}/(\bar{\ell} + K_\mathrm{D})$ with the dissociation constant $K_\mathrm{D} = k_-/k_+$, and the steady-state fraction of phosphorylated readout molecules is $f \equiv \bar{x}^*/X_\mathrm{T} = pR_\mathrm{T}/(pR_\mathrm{T} + k_\mathrm{r}/k_\mathrm{f})$.

In the linear-noise approximation (Appendix 3.B, [30]), expanding Eqs. 3.21 and 3.22 to first order around their steady state, the equations become

$$\delta\dot{RL} = b\delta\ell(t) - \delta RL(t)/\tau_\mathrm{c} + \eta_c(t), \tag{3.25}$$

$$\delta\dot{x}^* = \gamma\delta RL(t) - \delta x^*(t)/\tau_\mathrm{r} + \eta_x(t). \tag{3.26}$$

The parameters $b = R_\mathrm{T}p(1-p)/(\bar{\ell}\tau_\mathrm{c})$ and $\gamma = X_\mathrm{T}f(1-f)/(R_\mathrm{T}p\tau_\mathrm{r})$ are effective rates of receptor-ligand binding and readout phosphorylation, respectively. The decay rate of correlations in the receptor-ligand binding state is $\tau_\mathrm{c}^{-1} = \bar{\ell}k_+ + k_-$, and that of the readout phosphorylation state is $\tau_\mathrm{r}^{-1} = pR_\mathrm{T}k_\mathrm{f} + k_\mathrm{r}$. The rescaled white noise processes have strengths

$$\langle\eta_c^2\rangle = \bar{B}_c^2 = 2R_\mathrm{T}p(1-p)/\tau_\mathrm{c}, \tag{3.27}$$

$$\langle\eta_x^2\rangle = \bar{B}_x^2 = 2X_\mathrm{T}f(1-f)/\tau_\mathrm{r}. \tag{3.28}$$

MODEL STATISTICS

The required statistic to compute the past information is the variance in the output, decomposed into the part caused by signal variation and the part caused by noise. To compute the predictive information we further need the correlation function between the current output and a future ligand concentration $\langle\delta\ell(\tau)\delta x^*(0)\rangle$. These quantities can be obtained via their Fourier transforms, as in Appendix 3.C Eqs. 3.66 and 3.67. We can express the linear push-pull network as a 2-dimensional OU-process with $\boldsymbol{\delta y}(t) \equiv (\delta RL(t), \delta x^*(t))^T$,

$$\dot{\boldsymbol{\delta y}} = \boldsymbol{G}\delta\ell(t) + \boldsymbol{J\delta y}(t) + \boldsymbol{B\xi}(t), \tag{3.29}$$

also see Appendix 3.B. The matrices describing the properties of the signaling network are,

$$\boldsymbol{G} = \begin{pmatrix} b \\ 0 \end{pmatrix}, \tag{3.30}$$

$$\boldsymbol{J} = \begin{pmatrix} -\tau_\mathrm{c}^{-1} & 0 \\ \gamma & -\tau_\mathrm{r}^{-1} \end{pmatrix}, \tag{3.31}$$

$$\boldsymbol{B} = \begin{pmatrix} \sqrt{2R_\mathrm{T}p(1-p)/\tau_\mathrm{c}} & 0 \\ 0 & \sqrt{2X_\mathrm{T}f(1-f)/\tau_\mathrm{r}} \end{pmatrix}, \tag{3.32}$$

where the signal gain matrix $\boldsymbol{G}$ describes the strength by which the signal impacts each species directly, the Jacobian $\boldsymbol{J}$ of the signaling network describes its internal dynamics, and the matrix $\boldsymbol{B}$ gives the noise strengths.

A useful property of the network is the matrix exponential of its Jacobian, which in Fourier space is (see Appendix 3.C, Eqs. 3.63 and 3.65)

$$\begin{aligned} \mathcal{F}\{e^{\boldsymbol{J}t}\} &= (i\omega\mathbb{I}_2 - \boldsymbol{J})^{-1}, \\ &= \begin{pmatrix} \frac{1}{1/\tau_\mathrm{c}+i\omega} & 0 \\ \frac{\gamma}{(1/\tau_\mathrm{c}+i\omega)(1/\tau_\mathrm{r}+i\omega)} & \frac{1}{1/\tau_\mathrm{r}+i\omega} \end{pmatrix}. \end{aligned} \tag{3.33}$$

Together with the gain and noise matrices of Eqs. 3.30 and 3.32, this matrix specifies the frequency dependent gain matrix $\mathbb{K}(\omega) = \mathcal{F}\{e^{\boldsymbol{J}t}\}\boldsymbol{G}$, and the frequency dependent noise matrix $\mathbb{N}(\omega) = \mathcal{F}\{e^{\boldsymbol{J}t}\}\boldsymbol{B}$, see also Appendix 3.C.

The integration kernel that maps the ligand concentration onto the output of the push-pull network is given by the inverse Fourier transform of the second entry of $\mathbb{K}(\omega)$, which is the frequency dependent gain, $K_{\ell\to x}(\omega)$, from ℓ to x^*:

$$\begin{aligned} k(t) \equiv \mathcal{F}^{-1}\{K_{\ell\to x}(\omega)\} &= b\gamma\tau_\mathrm{c}\tau_\mathrm{r}\frac{1}{\tau_\mathrm{r}-\tau_\mathrm{c}}\left(e^{-t/\tau_\mathrm{r}} - e^{-t/\tau_\mathrm{c}}\right), \\ &= X_\mathrm{T}f(1-f)(1-p)/\bar{\ell}\frac{1}{\tau_\mathrm{r}-\tau_\mathrm{c}}\left(e^{-t/\tau_\mathrm{r}} - e^{-t/\tau_\mathrm{c}}\right), \end{aligned} \tag{3.34}$$

The so-called static gain of the network is the integral of this kernel over all time,

$$\bar{g}_{\ell\to x} = \int_0^\infty k(t)dt = X_\mathrm{T}f(1-f)(1-p)/\bar{\ell}. \tag{3.35}$$

This parameter quantifies how much a step change in the input concentration changes the steady-state level of the output: $\bar{g}_{\ell\to x} = \partial\bar{x}^*/\partial\bar{\ell}$. We will use this parameter in the statistical quantities that follow. The static gain is also given by $\bar{g}_{\ell\to x} = \bar{g}_{\ell\to RL}\bar{g}_{RL\to x}$, with $\bar{g}_{\ell\to RL} = p(1-p)R_{\mathrm{T}}/\bar{\ell}$ the static gain from $\bar{\ell}$ to RL and $\bar{g}_{RL\to x} = f(1-f)X_{\mathrm{T}}/(pR_{\mathrm{T}})$ the static gain from RL to x^*.

The matrix of power spectra of the signaling system is in general given by, if the input signal is uncorrelated to the intrinsic noise [31],

$$\mathbb{S}_y(\omega) = \mathbb{K}(-\omega)\mathbb{S}_s(\omega)\mathbb{K}(\omega)^T + |\mathbb{N}(\omega)|^2, \tag{3.36}$$

also see Appendix 3.C. The power spectrum of the Markovian ligand concentration (Eq. 3.3) is

$$S_\ell(\omega) = \langle|\delta\ell(\omega)|^2\rangle = \frac{2\sigma_\ell^2/\tau_\ell}{1/\tau_\ell^2+\omega^2}. \tag{3.37}$$

Substituting this signal power spectrum together with the frequency dependent gain matrix $\mathbb{K}(\omega) = \mathcal{F}\{e^{\boldsymbol{J}t}\}\boldsymbol{G}$ and the frequency dependent noise matrix $\mathbb{N}(\omega) = \mathcal{F}\{e^{\boldsymbol{J}t}\}\boldsymbol{B}$, computed using Eqs. 3.30, 3.32 and 3.33, in Eq. 3.36 yields the matrix of all network power spectra $\mathbb{S}_y(\omega)$. The power spectrum of the read-out is entry (2,2) of this matrix:

$$\begin{aligned} S_x(\omega) &= |K_{\ell\to x}(\omega)|^2 S_\ell(\omega) + |N_x(\omega)|^2 \\ &= \frac{2b^2\gamma^2\sigma_\ell^2/\tau_\ell}{(1/\tau_{\mathrm{r}}^2+\omega^2)(1/\tau_{\mathrm{c}}^2+\omega^2)(1/\tau_\ell^2+\omega^2)} + \frac{\gamma^2\langle\eta_c^2\rangle}{(1/\tau_{\mathrm{r}}^2+\omega^2)(1/\tau_{\mathrm{c}}^2+\omega^2)} + \frac{\langle\eta_x^2\rangle}{1/\tau_{\mathrm{r}}^2+\omega^2}, \end{aligned} \tag{3.38}$$

The variance in the read-out is the sum of variations caused by changes in the signal, and variations caused by noise (Eq. 3.5),

$$\sigma_x^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_x(\omega)d\omega = \sigma_{x|\eta}^2 + \sigma_{x|L}^2, \tag{3.39}$$

with

$$\begin{aligned} \sigma_{x|L}^2 &= X_{\mathrm{T}}f(1-f) + \bar{g}_{RL\to x}^2 R_{\mathrm{T}}p(1-p)\frac{1}{1+\tau_{\mathrm{r}}/\tau_{\mathrm{c}}}, \\ &= X_{\mathrm{T}}f(1-f)\left(1+\frac{X_{\mathrm{T}}f(1-f)(1-p)}{pR_{\mathrm{T}}(1+\tau_{\mathrm{r}}/\tau_{\mathrm{c}})}\right), \end{aligned} \tag{3.40}$$

$$\begin{aligned} \sigma_{x|\eta}^2 &= \frac{\bar{g}_{\ell\to x}^2\sigma_\ell^2}{(1+\tau_{\mathrm{c}}/\tau_\ell)(1+\tau_{\mathrm{r}}/\tau_\ell)}\left(1+\frac{\tau_{\mathrm{c}}\tau_{\mathrm{r}}}{\tau_\ell(\tau_{\mathrm{c}}+\tau_{\mathrm{r}})}\right), \\ &= \tilde{g}_{L\to x}^2\sigma_\ell^2, \end{aligned} \tag{3.41}$$

where $\bar{g}_{RL\to x} = \gamma\tau_r = X_{\mathrm{T}}f(1-f)/(R_{\mathrm{T}}p)$ is the static gain from the receptor to the readout, and the last line defines the *past* dynamic gain $\tilde{g}_{L\to x}$ which quantifies how strongly fluctuations in the full past trajectory propagate to the output. Equation 3.40 gives insight into the role of the different network components in shaping the noise in the readout. The first term on its right-hand side gives the contribution from the phosphorylation dynamics of the readout, $X_{\mathrm{T}}f(1-f)$, which cannot be averaged out. The second term is the contribution from the receptor switching noise. While the noise at the level of the receptor, $R_\mathrm{T}p(1-p)$, increases with the number of receptors, that propagated to the readout, $\bar{g}^2_{RL\to x}R_\mathrm{T}p(1-p)/(1+\tau_\mathrm{r}/\tau_\mathrm{c})$, decreases with R_T: increasing R_T allows for more instantaneous measurements, which decreases the sensing error. In addition, the propagation of the receptor noise to the output can be mitigated by increasing the integration time τ_r: this is the mechanism of time-averaging [15]. Equation 3.41 shows how variations in the ligand concentration contribute to the readout variance. Naturally, increasing the static gain $\bar{g}_{\ell\to x}$ (Eq. 3.35) amplifies all signal fluctuations. There exists also an integration time τ_r that maximizes the contribution from the past signal fluctuations as quantified by the dynamic gain $\tilde{g}_{L\to x}$: increasing τ_r means that fluctuations further back in time contribute to σ^2_x, yet increasing it too much, i.e. when $\tau_\mathrm{r} \gtrsim \tau_\ell$, means that the input dynamics will be averaged out.

PAST INFORMATION OF THE PUSH-PULL NETWORK

To compute the past information we use Eq. 3.6 with its corresponding SNR, using Eqs. 3.40 and 3.41:

$$\begin{aligned}\mathrm{SNR}^{-1}_\mathrm{past} &= \frac{\sigma^2_{x|L}}{\sigma^2_{x|\eta}}\\ &= \left(1+\frac{\tau_\mathrm{c}}{\tau_\ell}\right)\left(1+\frac{\tau_\mathrm{r}}{\tau_\ell}\right)\left(1+\frac{\tau_\mathrm{c}\tau_\mathrm{r}}{\tau_\ell(\tau_\mathrm{c}+\tau_\mathrm{r})}\right)^{-1}\left(\frac{(\bar{\ell}/\sigma_\ell)^2}{X_\mathrm{T}f(1-f)(1-p)^2}+\frac{(\bar{\ell}/\sigma_\ell)^2}{R_\mathrm{T}p(1-p)(1+\tau_\mathrm{r}/\tau_\mathrm{c})}\right).\end{aligned} \tag{3.42}$$

This expression has a clear interpretation. The error arising from the readout modification noise decreases with X_T because, while the modification noise itself increases with X_T, the squared gain $\bar{g}^2_{\ell\to x}$ goes as X^2_T (see Eq. 3.35). The error arising from receptor binding can be reduced by increasing R_T or τ_r. However, the integration time cannot be increased too much: The prefactor comes from the dynamic gain $\tilde{g}_{L\to x}$ for the past information, which determines how strongly signal fluctuations are amplified at the level of the readout, see Eq. 3.41. Increasing the gain helps to lift the signal above the receptor switching and readout modification noise. Yet, while for small τ_r the gain indeed increases with τ_r, it decreases with τ_r for $\tau_\mathrm{r} \gtrsim \tau_\ell$ as discussed above. The interplay between time averaging and gain gives rise to an optimal integration time that maximizes the past information. The steady state fraction of phosphorylated readouts that maximizes the past information is $f = 1/2$.

PREDICTIVE INFORMATION OF THE PUSH-PULL NETWORK

To determine the predictive information we need to compute the correlation function from the current output to the future ligand concentration $\langle\delta x(0)\delta\ell(\tau)\rangle$, as shown in Section 3.3, Eqs. 3.15 and 3.17. This requires the cross-spectrum from output to ligand concentration, which is given by (Eq. 3.67)

$$K_{\ell\to x}(-\omega)S_\ell(\omega) = \frac{b\gamma}{(1/\tau_\mathrm{c}-i\omega)(1/\tau_\mathrm{r}-i\omega)}\frac{2\sigma^2_\ell/\tau_\ell}{1/\tau^2_\ell+\omega^2}. \tag{3.43}$$

From this power spectrum we obtain the required correlation function by taking the inverse Fourier transform:

$$\begin{aligned}\langle \delta x(0)\delta \ell(\tau)\rangle &= \mathcal{F}^{-1}\{\tilde{g}_{\ell\to x}(-\omega)S_\ell(\omega)\},\\ &= \frac{\bar{g}_{\ell\to x}e^{-\tau/\tau_\ell}}{(1+\tau_{\mathrm{c}}/\tau_\ell)(1+\tau_{\mathrm{r}}/\tau_\ell)}\sigma_\ell^2 = \tilde{g}\sigma_\ell^2,\end{aligned} \tag{3.44}$$

which yields the dynamic gain $\tilde{g} \equiv \langle \delta x(0)\delta \ell(\tau)\rangle / \sigma_\ell^2$ (see Eq. 3.10):

$$\tilde{g} = \frac{\bar{g}_{\ell\to x}e^{-\tau/\tau_\ell}}{(1+\tau_{\mathrm{c}}/\tau_\ell)(1+\tau_{\mathrm{r}}/\tau_\ell)}, \tag{3.45}$$

with the static gain $\bar{g}_{\ell\to x}$ given by Eq. 3.35. Importantly, Eq. 3.44 reveals that the cross-correlation $\langle \delta x(0)\delta \ell(\tau)\rangle$ between the current output $x(0)$ and the future input $\ell(\tau)$ depends on the forecast interval τ via a factor that only depends on the signal correlation time τ_ℓ but not on the network design:

$$\langle \delta x(0)\delta \ell(\tau)\rangle = \langle \delta x(0)\delta \ell(0)\rangle e^{-\tau/\tau_\ell}. \tag{3.46}$$

This also means that the coefficient for the correlation between the current output and future input, $\rho_{x\ell}(\tau) \equiv \langle \delta x(0)\delta \ell(\tau)\rangle / (\sigma_x\sigma_\ell)$, is related to the instantaneous correlation coefficient $\rho_{x\ell}(0)$ via the same exponential factor:

$$\rho_{x\ell}(\tau) = \rho_{x\ell}(0)e^{-\tau/\tau_\ell}. \tag{3.47}$$

To compute the predictive information (see Eq. 3.17), we thus only need to specify the instantaneous correlation coefficient $\rho(0)$:

$$\rho_{x\to\ell}^{-2}(0) = \frac{\sigma_x^2\sigma_\ell^2}{\langle \delta x(0)\delta \ell(0)\rangle^2}. \tag{3.48}$$

$$\tag{3.49}$$

The variance σ_x^2 is given by Eqs. 3.39-3.41 and the instantaneous correlation $\langle \delta x(0)\delta \ell(0)\rangle^2$ is given by Eq. 3.44 with $\tau = 0$. Combining these results yields

$$\begin{aligned}\rho_{x\ell}^{-2}(0) = &\left(1+\frac{\tau_{\mathrm{c}}}{\tau_\ell}\right)^2\left(1+\frac{\tau_{\mathrm{r}}}{\tau_\ell}\right)^2\left(\frac{(\bar{\ell}/\sigma_\ell)^2}{X_{\mathrm{T}}f(1-f)(1-p)^2}+\frac{(\bar{\ell}/\sigma_\ell)^2}{R_{\mathrm{T}}p(1-p)(1+\tau_{\mathrm{r}}/\tau_{\mathrm{c}})}\right)\\ &+\left(1+\frac{\tau_{\mathrm{c}}}{\tau_\ell}\right)\left(1+\frac{\tau_{\mathrm{r}}}{\tau_\ell}\right)\left(1+\frac{\tau_{\mathrm{c}}\tau_{\mathrm{r}}}{\tau_\ell(\tau_{\mathrm{c}}+\tau_{\mathrm{r}})}\right).\end{aligned} \tag{3.50}$$

The predictive information then follows via Eq. 3.17. As for the past information, the steady state fraction of phosphorylated readouts that maximizes the predictive information is $f = 1/2$. As shown in Eq. 3.16, we note that the inverse SNR is related to the instantaneous correlation function as

$$\mathrm{SNR}_{\mathrm{pred}}^{-1} = \rho_{x\ell}^{-2}(0)e^{2\tau/\tau_\ell} - 1, \tag{3.51}$$

setting the predictive information via Eq. 3.11. We discuss the optimal design of the push-pull network that maximizes the predictive information in detail in Appendix 3.D.

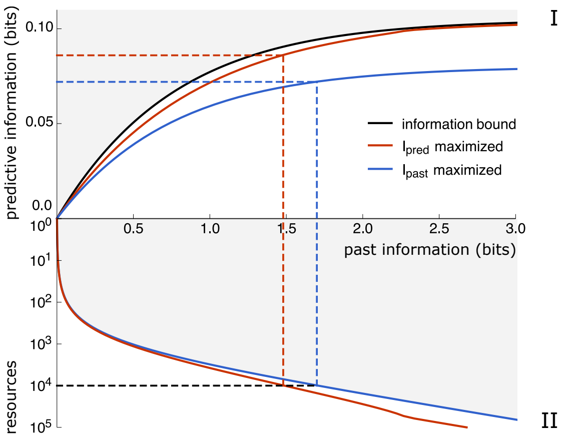


Figure 3.2: The optimal push-pull network is not at the information bound. Panel I: The black line is the information bound that maximizes the predictive information $I_{\text{pred}} = I(x_0; \ell_\tau)$ for a given past information $I_{\text{past}} = I(x_0; \boldsymbol{L}_p)$. The blue curve shows I_{pred} versus I_{past} for systems where I_{past} has been maximized for a given resource cost $C = R_\text{T} + X_\text{T}$. The red curve shows I_{pred} against I_{past} for systems in which I_{pred} has been maximized for a given C. Panel II shows I_{past} against C for the corresponding systems. The dashed lines indicate how the resource cost C sets I_{past} and hence I_{pred}. The forecast interval is $\tau = \tau_\ell$. The optimization parameters are the ratio X_T/R_T, τ_r, p and f (see Section 3.4 and Appendix 3.D). Other parameter values: $(\sigma_\ell/\bar{\ell})^2 = 10^{-2}$, $\tau_\text{c}/\tau_\ell = 10^{-2}$.

3.5. THE OPTIMAL PUSH-PULL NETWORK IS NOT AT THE INFORMATION BOUND

The topology of the push-pull network defines the shape of its integration kernel (Eq. 3.34), and thus which properties the network can extract from the signal in the past. How much information the cells can extract from the past signal however, depends on the resources devoted to building and operating the network (Figs. 3.1 and 3.2-II). We define the total resource cost as:

$$C = \lambda(R_\text{T} + X_\text{T}) + c_1 X_\text{T} \Delta\mu/\tau_\text{r}. \tag{3.52}$$

The first term expresses the fact that over the course of the cell cycle all components need to be duplicated, which means that they have to be synthesized at a speed that is at least the growth rate λ. The second term describes the chemical power that is necessary to run the push-pull network [13, 15]; it depends on the flux through the network, X_T/τ_r, and the free-energy drop $\Delta\mu$ over a cycle, e.g. the free energy of ATP hydrolysis in the case of a phosphorylation cycle. The coefficient c_1 describes the relative energetic cost of synthesizing the components during the cell cycle versus that of running the system. For simplicity, we first consider the scenario that the cost is dominated by that of protein synthesis, setting $c_1 \to 0$. While in this scenario $R_\text{T} + X_\text{T}$ is constrained, X_T/R_T and other system parameters are free for optimization.

The available resources C put a hard bound on the information I_{past} that can be extracted from the past signal, which in turn sets a hard limit on the predictive information

I_{pred} (Fig. 1.1C). However, this does not imply that the optimal system that maximizes the predictive information I_{pred} per resource cost C, $I_{\mathrm{pred}}/C = \left(I_{\mathrm{pred}}/I_{\mathrm{past}}\right)\left(I_{\mathrm{past}}/C\right)$, is also a maximally predictive system, maximizing $I_{\mathrm{pred}}/I_{\mathrm{past}}$, or a parsimonious system, maximizing I_{past}/C. To elucidate the interplay between I_{pred}, I_{past}, and C, we first maximize I_{past} for a given resource constraint C. We find a unique optimal design for the push-pull network (blue line in Fig. 3.2-II), which implies that not all bits of past information are equally costly. We then compute the corresponding predictive information for the systems along this line, which is the blue line in Fig. 3.2-I. Strikingly, the resulting information curve lies far below the information bound (black line, Fig. 3.2-I), demonstrating that parsimonious systems, which maximize I_{past}/C, are not maximally predictive. This is because not all bits of past information are equally predictive.

Precisely because bits of past information are neither equally predictive nor equally costly, it is paramount to directly maximize the predictive information for a given resource cost, I_{pred}/C, in order to obtain the most efficient prediction device. This yields the red lines in panels I and II in Fig. 3.2. It can be seen that compared to parsimonious systems (blue lines), the predictive information is higher while the past information is lower. While the bound on the predictive information as set by the resource cost (red line panel I) is close to the bound on the predictive information as set by the past information (black line), it does remain lower. This is surprising, because the push-pull network is a copying device [13, 47], which can, as we will also show below, reach the latter bound. Taken together, our results imply that the most cost-efficient prediction systems are neither parsimonious nor maximally predictive, but instead trade off those bits of past information that are most informative about the future, maximizing $I_{\mathrm{pred}}/I_{\mathrm{past}}$, against those that are cheapest, maximizing I_{past}/C.

In Appendix 3.D we discuss in detail the optimal design of the push-pull network in terms of its signal to noise ratio. Here, we continue by investigating how this optimal design arises from the trade-off between the cost and the predictive power of past information.

3.6. Trade-off between cost and predictive power per bit

To better understand the connection between predictive and past information, and resource cost, we map out the region in the information plane that can be reached given a resource constraint C (Fig. 3.3A, green region). We immediately make two observations. Firstly, the system can indeed reach the information bound. Secondly, the system can increase both the past and the predictive information by moving away from it. To elucidate these two observations, we investigate the system along the isocost line of $C = 10^4$, which together with the information bound envelopes the accessible region for the maximum resource cost $C \leq 10^4$.

Along the line $C = 10^4$, the ratio of the number of readout over receptor molecules is $X_{\mathrm{T}}/R_{\mathrm{T}} = 2\sqrt{p/(1-p)}\sqrt{1+\tau_{\mathrm{r}}/\tau_{\mathrm{c}}}$ (Appendix 3.D, Eq. 3.72). Systems that do not obey this relation are inside the accessible region (see also Appendix 3.D, Fig. 3.6), as are the systems with $C < 10^4$. The relation can be understood intuitively using the optimal resource allocation principle [13]. It states that in a sensing system that employs its proteins op-

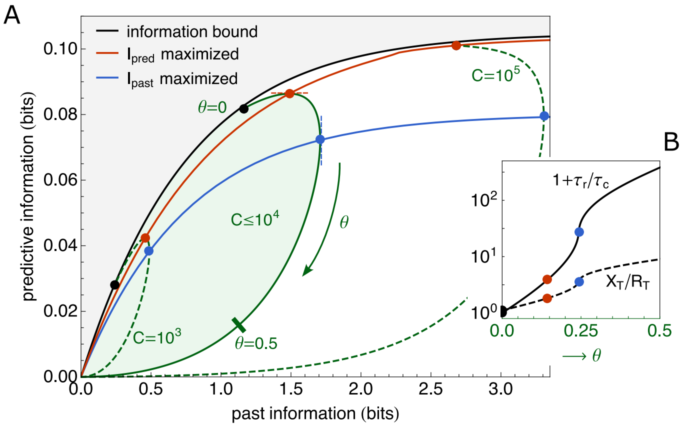


Figure 3.3: The push-pull network maximizes the predictive power under a resource constraint by moving away from the information bound. (A) The region of accessible predictive information $I_{\rm pred} = I(x_0;\ell_\tau)$ and past information $I_{\rm past} = I(x_0;\boldsymbol{L}_p)$ in the push-pull network under a resource constraint $C \leq R_{\rm T} + X_{\rm T}$, for the Markovian signals specified by Eq. 3.3 (green). The black line is the information bound at which $I_{\rm pred}$ is maximized for a given $I_{\rm past}$. The push-pull network can be at the information bound (black points), but maximizing $I_{\rm pred}$ for a resource constraint C moves the system away from it. The red and blue lines connect, respectively, the points where $I_{\rm pred}$ and $I_{\rm past}$ are maximized along the green isocost lines (the contourlines of constant C); they correspond to the red and blue lines in Fig. 3.2. The accessible region of $I_{\rm pred}$ and $I_{\rm past}$ for a given C has been obtained by optimizing over $\tau_{\rm r}$, p, f, and $X_{\rm T}/R_{\rm T}$. The forecast interval is $\tau = \tau_\ell$. (B) The integration time $\tau_{\rm r}$ over the receptor correlation time $\tau_{\rm c}$, $\tau_{\rm r}/\tau_{\rm c}$, and the ratio of the number of readout and receptor molecules, $X_{\rm T}/R_{\rm T}$, as a function of the distance θ along the isocost line corresponding to $C = 10^4$ in panel A; the red and blue points denote where $I_{\rm pred}$ and $I_{\rm past}$ are maximized along the contourline, respectively. For $\theta \to 0$, $\tau_{\rm r} \to 0$: the system is an instantaneous responder, and only the finite receptor correlation time $\tau_{\rm c}$ prevents the system from fully reaching the information bound; as predicted by the optimal resource allocation principle, $X_{\rm T} = R_{\rm T}$ [13]. The system can increase $I_{\rm pred}$ and $I_{\rm past}$ (A) by increasing $\tau_{\rm r}$ and $X_{\rm T}/R_{\rm T}$. Parameter values: $(\sigma_\ell/\bar{\ell})^2 = 10^{-2}$, $\tau_{\rm c}/\tau_\ell = 10^{-2}$.

timally, the total number of independent concentration measurements at the level of the receptor during the integration time $\tau_{\rm r}$, $R_{\rm T}(1+\tau_{\rm r}/\tau_{\rm c})$, equals the number of readout molecules $X_{\rm T}$ that store these measurements, so that neither the receptors nor the readout molecules are in excess. This design principle specifies, for a given integration time $\tau_{\rm r}$, the ratio $X_{\rm T}/R_{\rm T}$ at which the readout molecules sample each receptor molecule roughly once every receptor correlation time $\tau_{\rm c}$.

While the optimal allocation principle gives the optimal ratio $X_{\rm T}/R_{\rm T}$ of the number of readouts over receptors for a *given* integration time $\tau_{\rm r}$, it does not prescribe what the optimal integration time $\tau_{\rm r}^{\rm opt}$, and hence the (globally) optimal ratio $(X_{\rm T}/R_{\rm T})^{\rm opt}$, is that maximizes $I_{\rm pred}$ for a given resource constraint $C = R_{\rm T} + X_{\rm T}$. Figure 3.3B shows that as the distance θ along the isocost line is increased, $\tau_{\rm r}$ and hence $X_{\rm T}/R_{\rm T}$ increase monotonically. For $\theta \to 0$, the integration time $\tau_{\rm r}$ is zero and the number of readout molecules equals the number of receptor molecules: $X_{\rm T} = R_{\rm T}$. In this limit, the push-pull network

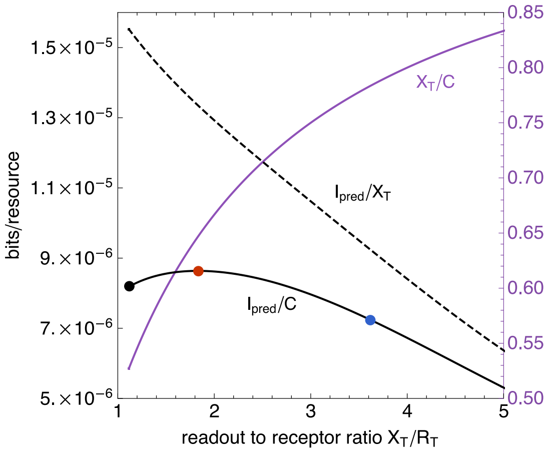


Figure 3.4: The optimal design arises from a trade-off between the cost and the predictive information per physical bit. Predictive information per physical bit (dashed line), physical bits per resource cost (purple line), and predictive information per resource cost (solid black line), as a function of X_T/R_T parametrically along the contourline $C = 10^4$ of Fig. 3.3A (such that the optimal allocation ratio, Eq. 3.72, is obeyed). The colored dots on the solid black curve correspond to those along the contourline $C = 10^4$ in Fig. 3.3A. While increasing the ratio of readout to receptor molecules X_T/R_T and the integration time τ_r decreases the predictive information I_{pred} per physical bit of past information, I_{pred}/X_T, increasing X_T/R_T does increase the number of physical bits per resource cost, X_T/C. This trade-off gives rise to an optimal predictive information per resource cost, I_{pred}/C (red dot). Parameter values as in Fig. 3.3.

is an instantaneous responder, with an integration kernel given by Eq. 3.4. The system is indeed very close to the information bound; only the finite receptor correlation time τ_c prevents the system from fully reaching it. Yet, as θ increases and the system moves away from the bound, the predictive and past information first rise along the contour, and thus with X_T/R_T and τ_r, before they eventually both fall.

To understand why the predictive and past information first rise and then fall with X_T/R_T and τ_r, we note that each readout molecule constitutes 1 physical bit and that its binary state (phosphorylated or not) encodes at most 1 bit of information on the ligand concentration. The number of readout molecules X_T thus sets a hard upper bound on the sensing precision and hence the predictive information. To raise this bound, X_T must be increased. For a given resource constraint $C = R_T + X_T$, X_T can only be increased if the number of receptors R_T is simultaneously decreased. However, the cell infers the concentration not from the readout molecules directly, but via the receptor molecules: a readout molecule is a sample of the receptor that provides at most 1 bit of information about the ligand-binding state of a receptor molecule, which in turn provides at most 1 bit of information about the input signal. To raise the lower bound on the predictive information, the information on the input must increase at both the receptor and the readout level.

To elucidate how this can be achieved, we note that the maximum number of independent receptor samples and hence concentration measurements is given by $N_{\rm I}^{\rm max} = \min(X_{\rm T}, R_{\rm T}(1+\tau_{\rm r}/\tau_{\rm c}))$ [13]. For $\theta > 0$, the system can increase $N_{\rm I}^{\rm max}$ if, and only if, $X_{\rm T}$ and $R_{\rm T}(1+\tau_{\rm r}/\tau_{\rm c})$ can be raised simultaneously. This can be achieved, while obeying the constraint $C = X_{\rm T} + R_{\rm T}$, by decreasing $R_{\rm T}$ yet increasing $\tau_{\rm r}$ (Fig. 3.3B). This is the mechanism of time averaging, which makes it possible to increase the number of independent receptor samples [15], and explains why both $I_{\rm pred}$ and $I_{\rm past}$ initially increase (Fig. 3.3A, Fig. 3.4). However, as $\tau_{\rm r}$ is raised further, the receptor samples become older: the readout molecules increasingly reflect receptor states in the past that are less informative about the future ligand concentration. The collected bits of past information have become less predictive about the future (Fig. 3.4). For a given resource cost, the cell thus faces a trade-off between maximizing the number of physical bits of past information (i.e. receptor samples $X_{\rm T}$) and the predictive information per bit. This antagonism gives rise to the optimal integration time $\tau_{\rm r}^{\rm opt}$ and ratio $(X_{\rm T}/R_{\rm T})^{\rm opt}$ (Fig. 3.3A/B) that maximizes the total predictive information $I_{\rm pred}$ (Fig. 3.4). As C is increased the system moves towards the information bound (see Fig. 3.3A) because the relative cost of physical bits decreases, and the balance therefore shifts towards maximizing the predictive information per bit, yielding a smaller $\tau_{\rm r}^{\rm opt}$.

Interestingly, while $I_{\rm pred}$ decreases beyond $\tau_{\rm r}^{\rm opt}$, the past information $I_{\rm past}$ first continues to rise because $N_{\rm I}^{\rm max}$ still increases (Fig. 3.3A). However, when the integration time becomes longer than the input signal correlation time, the correlation between input and output is lost and $I_{\rm past}$ falls too.

3.7. CHEMICAL POWER PREVENTS REACHING THE INFORMATION BOUND

So far, we have only considered the cost of maintaining the cellular system, the protein cost $C = R_{\rm T} + X_{\rm T}$. Yet, driving a push-pull network also requires energy. In fact, for *E. coli* at a typical cell doubling time $\lambda^{-1} \simeq 1$hr, the operating cost is comparable to the maintenance cost [13, 49]. As Eq. 3.52 shows, the operating cost scales with the flux around the phosphorylation cycle, which is proportional to the inverse of the integration time, $\tau_{\rm r}^{-1}$. The power thus diverges for $\tau_{\rm r} \to 0$. Since the information bound is reached precisely in this limit, it is clear that the chemical power prevents the push-pull network from reaching the bound (see Fig. 3.5A). Indeed, when we consider the operating costs, the push-pull network can only be at the information bound when $(I_{\rm past}, I_{\rm pred}) \to (0,0)$ or $C \to \infty$ (Fig. 3.5A). The system can mitigate the operating costs by decreasing $X_{\rm T}$, because this decreases the flux through the cycle. However, this also decreases the gain and thus, eventually, any information transduced through the network. In the limit that both $X_{\rm T}$ and $\tau_{\rm r}$ approach zero, the system approaches the information bound at the origin (Fig. 3.5A and B). More generally, when the operating costs are taken into account, the system time averages more (i.e., $\tau_{\rm r}$ rises), because frequent measurements are now even more costly. Still, $\tau_{\rm r}$ decreases as the total resource availability C grows.

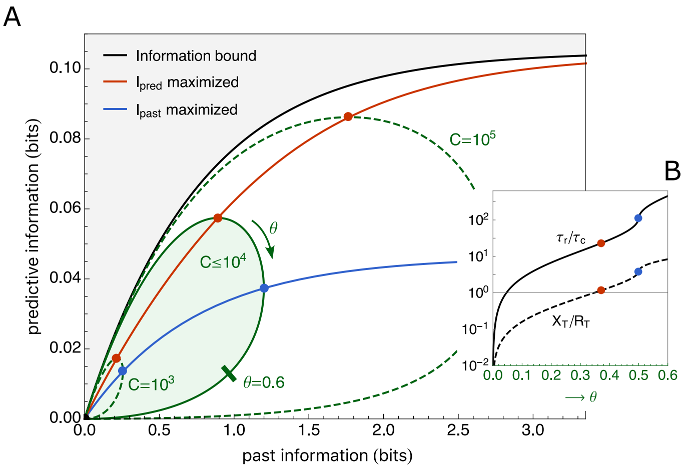


Figure 3.5: The cost of operating the network moves the system away from the information bound. (A) The region of accessible predictive information $I_{\text{pred}} = I(x_0;\ell_\tau)$ and past information $I_{\text{past}} = I(x_0;\boldsymbol{L}_p)$ in the push-pull network under a resource constraint that is not only given by a protein cost but also an operating cost, $C \leq \lambda(R_T + X_T) + c_1 X_T \Delta\mu/\tau_r$, for the Markovian signals specified by Eq. 3.3 (green). The black line is the information bound at which I_{pred} is maximized for a given I_{past}. The cost of operating the network, $\sim X_T\Delta\mu/\tau_r$, moves the system away from the information bound (compare to Fig. 3.3). The red and blue lines correspond to systems that maximize I_{pred} and I_{past}, respectively, under the constraint C. The accessible region of I_{pred} and I_{past} for a given C has been obtained by optimizing over τ_r, p, f, and X_T/R_T. (B) The integration time over the receptor correlation time, τ_r/τ_c, and the ratio of the number of readout and receptor molecules, X_T/R_T, as a function of the distance θ along the isocost line for $C = 10^4$ in panel A. For $\theta \to 0$, both τ_r and X_T go to zero, thus reducing both I_{past} and I_{pred} to zero. Parameter values: $(\sigma_\ell/\bar{\ell})^2 = 10^{-2}$, $\tau_c/\tau_\ell = 10^{-2}$, $\tau = \tau_\ell$, $\lambda^{-1} = 1\text{h}, c_1 = 10^{-4}/\Delta\mu$. The estimate of c_1 is based on the operating and maintenance cost of the push-pull network of the *E. coli* chemotaxis system [13, 49]: one full cycle of readout (CheY) phosphorylation and dephosphorylation requires (at least) 1 ATP, while the synthesis of a readout or receptor (Tar/Tsr/Trg) protein requires about 10^4 ATPs; this estimate includes the cost of synthesizing the phosphatase CheZ, the kinase CheA and the adapter protein CheW. With τ_c = 10ms, τ_r of the optimal system is roughly 100ms, as also observed experimentally for the *E. coli* chemotaxis system [13, 50]. At a cell doubling time of approximately $\lambda^{-1} \simeq 1$h, and with $\tau_r = 10^{-4}$h, the maintenance cost is of the same order of magnitude as the operating cost [13]. The units of C are chosen such that they correspond to the number of proteins per hour.

3.8. DISCUSSION

Cellular systems need to predict the future signal by capitalizing on information that is contained in the past signal. To this end, they encode the past signal into the dynamics of an intracellular biochemical network from which the future input is inferred. To maximize the predictive information for a given amount of information that is extracted from the past, the cell should store those signal characteristics that are most informative about the future signal. For a Markovian signal this is the current signal value. As we have seen here, the push-pull network can copy the current input into the output. Therefore, this biochemical signaling network system is in principle able to extract the most predictive information from the past, allowing it to reach the information bound. Yet, our analysis also shows that extracting the most relevant information can be exceedingly costly in terms of biophysical resources.

So far, bounds on predictive information have been studied using information compression constraints [10–12]. Yet, while information is a resource—the cell cannot predict the future without extracting information from the past signal—the principal resources that have a direct cost are time, building blocks and energy. The predictive information per unit protein and energy cost is therefore most likely a more relevant fitness measure than the predictive information per bit past information.

Our analysis reveals that it is not always optimal to operate at the information bound: cells can increase the predictive information for a given resource constraint by moving away from the bound. Increasing the integration time in the push-pull network firstly reduces the chemical power, and secondly makes it possible to take more concentration measurements per protein copy, enabling the network to extract more information from the past signal. Yet, increasing the integration time also means that the information that has been collected, is less informative about the future signal. This interplay gives rise to an optimal integration time, which maximizes the predictive information for a given resource constraint. This argument also explains why the system moves towards the information bound when the resource constraint is relaxed: increasing the number of receptor and readout molecules allows the system to take more instantaneous concentration measurements, which makes time averaging less important, thus reducing the integration time.

We have determined how much predictive information the push-pull network can collect under varying resource availability. Perhaps surprisingly, the amount of predictive information the network can collect as measured in bits seems exceedingly small. For a resource availability of $C = 10^4$ the network obtains less then one tenth of a bit predictive information, a mere 5% of the information extracted from the past (Figs. 3.3 and 3.5). However, these numbers strongly depend on the forecast interval τ. Throughout this work we consider a forecast interval equal to the signal correlation time, which sets a maximal reasonable interval for prediction, and therefore yields an effective minimum on the predictive information. In reality, we expect the relevant timescale setting the forecast interval to be the time that it takes the cell to mount a response, which must be shorter than the signal correlation time for a successful predict-and-anticipate strategy. For the Markovian signal considered here, the optimal design of the signaling network is independent of the prediction interval. Yet, for a resource availability $C = 10^4$, the predictive information increases up to almost one bit as the forecast interval decreases to zero. In this regime the network can use almost 80% of the information it extracts from the past for prediction. Regardless of the prediction interval, the optimal system is close to, but not at the information bound.

This chapter vividly demonstrates that not all bits of past information are equally predictive, nor costly. Specifically, the most recent bits of information are the most predictive, but also the most costly. While we have shown this for a push-pull network sensing a Markovian input signal, it is reasonable to expect that for many classes of input signal the most recent information is the most predictive. Moreover, it is clear that it is costly to process signals both accurately and rapidly for a wide range of systems. To support this intuition, in the next chapter we turn our attention to a simple class of non-Markovian signals, and investigate how close one of the most well-studied biochemical signaling networks is to the information bound.

Appendices

3.A. Langevin equation

Throughout this dissertation we use the chemical Langevin notation for stochastic processes, and we approximate all biochemical networks as linear. In this appendix we first shortly recapitulate the rationale behind the Langevin notation (closely following [51]). In Appendix 3.B we then elaborate on the linear noise approximation [30].

Let the vector $\boldsymbol{Y}_t = \left(y_1(t), y_2(t), \ldots, y_n(t)\right)^T$ describe the state of the system, where $y_i(t)$ is the number of molecules of the i^{th} species at time t. Species can be involved in any of the m reactions R_j, the number of species i that is used or created in reaction j is given by s_{ij}, which is an entry of the stoichiometry matrix. The change in the number of species i over a time τ is then generally given by

$$y_i(t+\tau) = y_i(t) + \sum_{j=1}^{m} s_{ij} K_j(\boldsymbol{Y}_t, \tau), \tag{3.53}$$

where $K_j(\boldsymbol{Y}_t, \tau)$ is the number of times that reaction R_j occurs in the time interval $[t, t+\tau]$, which is a random variable. Obtaining $K_j(\boldsymbol{Y}_t, \tau)$ exactly would involve solving the chemical master equation, but we can approximate it in a convenient manner by imposing two conditions on the time interval τ.

Before defining these conditions we introduce the propensity function $a_j(\boldsymbol{Y}_t)$. This function gives the probability that reaction R_j occurs when the system is in state $\boldsymbol{Y}_t$ per unit time, and it thus depends on time via $\boldsymbol{Y}_t$. Such functions are well-defined in homogeneous systems where diffusion is fast relative to the time-scale of the reactions.

The first condition to approximate $K_j(\boldsymbol{Y}_t, \tau)$ is as follows: *We require τ to be small enough such that the change in system state during $[t, t+\tau]$ is so slight that the propensity functions remain approximately constant.* The propensity functions then satisfy

$$a_j(\boldsymbol{Y}_t) \cong a_j(\boldsymbol{Y}_{t'}), \quad \forall t' \in [t, t+\tau], \quad \forall j. \tag{3.54}$$

Apart from making τ very small, this condition is easier to obey when the system contains many molecules. The number of times that reaction R_j occurs in the time interval $[t, t+\tau]$ is now an independent Poisson distributed random variable with mean and variance $a_j(\boldsymbol{Y}_t)\tau$, i.e. $K_j(\boldsymbol{Y}_t, \tau) = \mathcal{P}_j(a_j(\boldsymbol{Y}_t), \tau)$.

To simplify this further we impose the second condition on τ: *We require τ to be large enough such that the expected number of occurrences of each reaction R_j in $[t, t+\tau]$ is much larger than 1, i.e.*

$$\langle \mathcal{P}_j(a_j(\boldsymbol{Y}_t), \tau) \rangle = a_j(\boldsymbol{Y}_t)\tau \gg 1, \quad \forall j. \tag{3.55}$$

This condition opposes the first condition, and it could be that both cannot be satisfied simultaneously. However, like the first condition, the second condition is helped by high molecule numbers in the system. We can now approximate the Poisson random variable by a normal random variable with equal mean and variance, $\mathcal{N}_j(a_j(\boldsymbol{Y}_t)\tau, a_j(\boldsymbol{Y}_t)\tau)$. Using that $\mathcal{N}(\mu, \sigma^2) = \mu + \sigma\mathcal{N}(0,1)$ we rewrite (3.53) as

$$y_i(t+\tau) - y_i(t) = \sum_{j=1}^{m} s_{ij} a_j(\boldsymbol{Y}_t)\tau + \sum_{j=1}^{m} s_{ij} \sqrt{a_j(\boldsymbol{Y}_t)} \mathcal{N}_j(0, \tau). \tag{3.56}$$

The deterministic part of the network and the diffusive effect of the noise are now separated in the two terms on the right-hand side, where the noise is caused by random variable $\mathcal{N}(0,\tau)$. This Gaussian random variable of mean 0 and variance τ is generated by a Brownian motion, or equivalently Wiener process, of duration τ. Finally, we define our timestep τ as a 'macroscopic infinitesimal' and denote it dt. Dividing both sides by dt then gives the Langevin notation

$$\dot{y}_i = \sum_{j=1}^{m} s_{ij} a_j(\boldsymbol{Y}_t) + \sum_{j=1}^{m} s_{ij}\sqrt{a_j(\boldsymbol{Y}_t)}\xi_j(t), \tag{3.57}$$

where each of the $\xi_j(t)$ is an independent Gaussian unit white noise process, which has mean 0 and is delta correlated $\langle \xi_i(t)\xi_j(t')\rangle = \delta_{ij}\delta(t-t')$, with δ_{ij} the Kronecker delta. These white noise processes arise as a convention for the derivative of Brownian motion. We make two final adjustments for ease of notation. Firstly, we define $f_i(\boldsymbol{Y}_t) \equiv \sum_{j=1}^{m} s_{ij} a_j(\boldsymbol{Y}_t)$ for the deterministic dynamics. Secondly, we write the diffusive term as a product of the $1 \times m$ vector of diffusion prefactors $\boldsymbol{B}_i(\boldsymbol{Y}_t)$ (which is the i^{th} row of diffusion matrix $\boldsymbol{B}(\boldsymbol{Y}_t)$), and the $m \times 1$ vector of independent white noise processes $\boldsymbol{\xi}(t)$. We then obtain the general Langevin form,

$$\dot{y}_i = f_i(\boldsymbol{Y}_t) + \boldsymbol{B}_i(\boldsymbol{Y}_t)\boldsymbol{\xi}(t). \tag{3.58}$$

SUMMARIZING NOISE TERMS

In some cases it is convenient to reduce the dimensions of the diffusion matrix by redefining the noise terms. This is specifically useful when each reaction only contributes to the change in molecule number of one species of interest [31]. For signal transduction networks in general this is often the case, and it also holds for the networks studied in this dissertation. In general the noise in any of the species is a sum of m independent white noise processes, (3.57), where m is the number of reactions. This sum has a cumulative mean of 0 and variance

$$\left\langle \sum_{j=1}^{m} s_{ij}\sqrt{a_j(\boldsymbol{Y}_t)}\xi_j(t) \sum_{k=1}^{m} s_{ik}\sqrt{a_k(\boldsymbol{Y}_t)}\xi_k(t') \right\rangle = \sum_{j=1}^{m} s_{ij}^2 a_j(\boldsymbol{Y}_t)\delta(t-t'), \tag{3.59}$$

because all m noise processes are independent. We now collect the independent Gaussian white noise processes that contribute to changes in the i^{th} species to define:

$$\boldsymbol{B}_i(\boldsymbol{Y}_t)\xi_i(t) \equiv \sum_{j=1}^{m} s_{ij}\sqrt{a_j(\boldsymbol{Y}_t)}\xi_j(t), \tag{3.60}$$

where we must have that $\boldsymbol{B}_i(\boldsymbol{Y}_t) = \sqrt{\sum_{j=1}^{m} s_{ij}^2 a_j(\boldsymbol{Y}_t)}$, by (3.59). We thus obtain n cumulative white noise processes, where n is the number of species, which are only independent if none of the reactions change the molecule numbers of multiple species. The alternative diffusion matrix $\boldsymbol{B}$ then becomes an $n \times n$ diagonal matrix.

3.B. LINEAR NOISE APPROXIMATION

Since the studied systems will generally have a single steady state, and we do not expect the noise to qualitatively change the mean dynamics, we further simplify the system dynamics using the classical linear noise approximation. We define deviations from the steady state as $\delta y_i(t) = y_i(t) - \bar{y}_i$. The steady state $\bar{y}_i$ is found by solving $f_i(\bar{\boldsymbol{y}}) = 0$. A Taylor expansion of (3.58) around $\bar{y}_i$, up to first order in the deterministic dynamics and to zeroth order in the noise, then gives

$$\delta \dot{y}_i = \sum_{k=1}^{n} \delta y_k(t) \left. \frac{\partial f_i(\boldsymbol{Y}_t)}{\partial y_k} \right|_{\bar{\boldsymbol{y}}} + \boldsymbol{B}_i \boldsymbol{\xi}(t). \tag{3.61}$$

The time derivative of $\bar{y}_i$ and the deterministic dynamics at steady state $f_i(\bar{\boldsymbol{y}})$ have dropped out because they are 0 by definition. We approximate the noise only to zeroth order because its dynamics are on the timescale $\sqrt{dt}$, which is much larger than dt. This can be most easily seen from (3.56) using $\mathcal{N}_j(0,\tau) = \sqrt{\tau}\mathcal{N}_j(0,1)$. For ease of notation we have defined $\boldsymbol{B}_i(\bar{\boldsymbol{y}}) \equiv \boldsymbol{B}_i$, which has entries $s_{ij}\sqrt{a_j(\bar{\boldsymbol{y}})}$ with $j = 1,\ldots,m$. We can further recognize that $\left.\frac{\partial f_i(\boldsymbol{Y}_t)}{\partial y_k}\right|_{\bar{\boldsymbol{y}}} = \boldsymbol{J}_{ik}$, i.e. the $(i,k)^{\text{th}}$ entry of the $n \times n$ Jacobian matrix $\boldsymbol{J}$ of the system.

We can express the dynamics of any linear signaling network as an n-dimensional OU-process

$$\dot{\boldsymbol{\delta y}} = \boldsymbol{G}\boldsymbol{\delta s}(t) + \boldsymbol{J}\boldsymbol{\delta y}(t) + \boldsymbol{B}\boldsymbol{\xi}(t), \tag{3.62}$$

where $\boldsymbol{\delta s}(t)$ is a $k \times 1$ vector of input signals and $\boldsymbol{\delta y}$ is the $n \times 1$ vector of network species, both defined in terms of deviations from their mean. The $n \times k$ signal gain matrix $\boldsymbol{G}$ describes the strength by which each signal impacts each species directly, the $n \times m$ matrix $\boldsymbol{B}$ contains the noise strengths. The eigenvalues of the Jacobian $\boldsymbol{J}$ must be negative for the system to be stable, and we require all signals to be stationary.

The form of Eq. 3.62, where we have explicitly written the signals as external to the system, will be convenient when deriving the general forms of the network power spectra and correlation functions in Appendix 3.C.

3.C. POWER SPECTRA, KERNELS, CORRELATION FUNCTIONS

In this appendix we derive the stationary auto-correlation matrix of a multidimensional OU-process, such as Eq. 3.62, via the networks' power spectra. The power spectrum of a real-valued random process $X(t)$ is the squared modulus of its Fourier transform: $S_x(\omega) = \langle \delta\tilde{x}(-\omega)\delta\tilde{x}(\omega)\rangle$ and $S_{x\to y}(\omega) = \langle \delta\tilde{x}(-\omega)\delta\tilde{y}(\omega)\rangle$. Throughout this work we use the following conventions for the Fourier transform and its inverse: $\mathcal{F}\{f(t)\} \equiv \tilde{f}(\omega) = \int_{-\infty}^{\infty} dt f(t)\exp(-i\omega t)$ and $\mathcal{F}^{-1}\{\tilde{f}(\omega)\} = 1/(2\pi)\int_{-\infty}^{\infty} d\omega \tilde{f}(\omega)\exp(i\omega t) = f(t)$. To obtain the correlation functions from the power spectra we invoke the Wiener-Khinchin theorem.

The general solution to Eq. 3.62 is

$$\boldsymbol{\delta y}(t) = \int_{-\infty}^{t} dt' \, e^{\boldsymbol{J}(t-t')} \left(\boldsymbol{G}\boldsymbol{\delta s}(t') + \boldsymbol{B}\boldsymbol{\xi}(t')\right), \tag{3.63}$$

which shows the two contributions to the time dependent solution: that of the external signal and that of the internal noise. The $n \times k$ matrix $e^{\boldsymbol{J}(t-t')}\boldsymbol{G}$ contains the integration kernels, its $(i,j)^{\text{th}}$ entry determines how the j^{th} signal affects the i^{th} system component over time. The $n \times m$ matrix $e^{\boldsymbol{J}(t-t')}\boldsymbol{B}$ is similar, but contains the functions that map the noise terms onto the system components. These matrices can be obtained by taking the

Fourier transform of Eq. 3.62 and solving for $\boldsymbol{\delta\tilde{y}}(\omega)$

$$i\omega\boldsymbol{\delta\tilde{y}}(\omega) = \boldsymbol{G\delta\tilde{s}}(\omega) + \boldsymbol{J\delta\tilde{y}}(\omega) + \boldsymbol{B\tilde{\xi}}(\omega), \tag{3.64}$$

$$\boldsymbol{\delta\tilde{y}}(\omega) = (i\omega\mathbb{I}_n - \boldsymbol{J})^{-1}\left(\boldsymbol{G\delta\tilde{s}}(\omega) + \boldsymbol{B\tilde{\xi}}(\omega)\right). \tag{3.65}$$

Using the convolution theorem to take the Fourier transform of Eq. 3.63, and comparing the result to Eq. 3.65, shows that $\mathcal{F}\{e^{\boldsymbol{J}(t-t')}\} = (i\omega\mathbb{I}_n - \boldsymbol{J})^{-1}$. We obtain for the power-spectra of the network components

$$\begin{aligned}\mathbb{S}_y(\omega) &= \langle\delta\tilde{\boldsymbol{y}}(-\omega)\delta\tilde{\boldsymbol{y}}(\omega)^T\rangle,\\ &= \mathbb{K}(-\omega)\mathbb{S}_s(\omega)\mathbb{K}(\omega)^T + |\mathbb{N}(\omega)|^2,\end{aligned} \tag{3.66}$$

with the matrices of frequency dependent gains $\mathbb{K}(\omega) \equiv (i\omega\mathbb{I}_n - \boldsymbol{J})^{-1}\boldsymbol{G}$, and frequency dependent noise $\mathbb{N}(\omega) \equiv (i\omega\mathbb{I}_n - \boldsymbol{J})^{-1}\boldsymbol{B}$. The cross terms vanish because we assume the fluctuations of the external signal are uncorrelated with the internal network noise. Furthermore, the power spectrum of a white noise process is constant, and all the noise terms are independent of one another, such that the spectral density of the noise vector is the identity matrix $\langle\tilde{\boldsymbol{\xi}}(-\omega)\tilde{\boldsymbol{\xi}}(\omega)^T\rangle = \mathbb{I}_m$. We also need to consider the cross-spectra between the signals and the network components, specifically we will need the spectra from the network to the signals

$$\begin{aligned}\mathbb{S}_{y\to s}(\omega) &= \langle\delta\tilde{\boldsymbol{y}}(-\omega)\delta\tilde{\boldsymbol{s}}(\omega)^T\rangle,\\ &= \mathbb{K}(-\omega)\mathbb{S}_s(\omega).\end{aligned} \tag{3.67}$$

From Eq. 3.66 and Eq. 3.67 we can obtain all necessary correlation functions and (co-)variances, by taking the inverse Fourier transform of the component of interest (for a variance we can directly set $t = 0$). The advantage of using this form, is that the contribution of each signal and of the noise terms appear separately. When we are for example interested in a variance that is only caused by noise, we can omit the terms depending on the signal power spectra, and vice versa. Moreover, the power spectra often have a simpler analytical form than the corresponding correlation functions.

3.D. OPTIMAL DESIGN VIA THE SIGNAL TO NOISE RATIO

To optimally predict the future, the cell must lift the signal above the molecular noise, and extract those signal features that are most informative about the future [13, 15, 29]. This is reflected in the signal-to-noise ratio (SNR) that sets the predictive information, Eqs. 3.7-3.13 and Eqs. 3.50 and 3.51. Here, we will discuss how the network gain and noise that make up the relative prediction error $\text{SNR}^{-1}_{\text{pred}}$, are affected by the number of available resources (Eq. 3.52) and the integration time τ_r. This will elucidate the optimal design of the network that maximizes the predictive information under constrained resource availability.

Optimal design is independent of forecast interval. While in general the optimal prediction strategy might depend on the forecast interval τ, Eq. 3.47 in combination with Eq. 3.17 shows that in this system the forecast interval only affects the value of the predictive information: the optimal design of the network that maximizes the predictive information does not depend on τ, because the signal is Markovian—the design that optimally estimates the current signal, which contains all the information on the future signal, therefore also equals the design that optimally predicts the future signal. The optimal design is obtained by minimizing the instantaneous inverse correlation coefficient, $\rho_{x\ell}^{-2}(0)$, given by Eq. 3.50.

Noise. The first part of the signal-to-noise ratio that we will discuss is the noise $\mathrm{NOISE}_{\mathrm{pred}} = \sigma^2_{x|\ell_\tau}$ (see Eq. 3.13), which is given by $\sigma^2_{x|\ell_\tau} = \sigma^2_x - \tilde{g}^2\sigma^2_\ell$, see Eq. 3.9. With the total variance σ^2_x given by Eqs. 3.39-3.41 and the dynamic gain $\tilde{g}$ given by Eq. 3.45, this yields:

$$\mathrm{NOISE}_{\mathrm{pred}} = \sigma^2_{x|L} + \sigma^2_{x|\eta} - \tilde{g}^2\sigma^2_\ell \tag{3.68}$$

$$\begin{aligned} &= X_{\mathrm{T}} f(1-f) + \bar{g}^2_{RL\to x}\frac{1}{1+\tau_{\mathrm{r}}/\tau_{\mathrm{c}}} R_{\mathrm{T}} p(1-p) + \\ &\quad \tilde{g}^2\sigma^2_\ell \left(e^{2\tau/\tau_\ell}\left(1+\frac{\tau_{\mathrm{c}}}{\tau_\ell}\right)\left(1+\frac{\tau_{\mathrm{r}}}{\tau_\ell}\right)\left(1+\frac{\tau_{\mathrm{c}}\tau_{\mathrm{r}}}{\tau_\ell(\tau_{\mathrm{c}}+\tau_{\mathrm{r}})}\right) - 1\right). \end{aligned} \tag{3.69}$$

This expression reveals that the total noise has three contributions: (1) readout modification; (2) receptor switching; (3) signal variations. The readout (de)modification noise is $X_{\mathrm{T}} f(1-f)$; since this is the noise added in the final step of signal propagation from $\ell \to x^*$, it cannot be averaged out. The receptor switching noise at the level of the receptors is $R_{\mathrm{T}} p(1-p)$, but that propagated to the output x^* is the time-averaged amount $\bar{g}^2_{RL\to x}/(1+\tau_{\mathrm{r}}/\tau_{\mathrm{c}})R_{\mathrm{T}} p(1-p)$. This contribution decreases by increasing τ_{r} via the mechanism of time averaging. Since the static gain from the receptor to the readout $\bar{g}_{RL\to x} = X_{\mathrm{T}} f(1-f)/(R_{\mathrm{T}} p)$ decreases with R_{T} because at high R_{T} the signal saturates, the receptor-switching noise contribution to σ^2_x decreases with R_{T}, scaling as $1/R_{\mathrm{T}}$. The third contribution, from the signal variations, depends non-trivially on the integration time τ_{r}: there exists a τ_{r} that minimizes this contribution.

Gain and Signal. The signal-to-noise ratio $\mathrm{SNR}_{\mathrm{pred}}$ depends not only on the noise, but also on the dynamic gain $\tilde{g}$, which together with the input variance σ^2_ℓ sets the signal, see Eq. 3.12. The dynamic gain $\tilde{g}$, given by Eq. 3.45, quantifies the degree to which variations in the output x^* are correlated with the (future) input. The dynamic gain depends on the static gain $\bar{g}_{\ell\to x}$, which is the product of the static gain $\bar{g}_{\ell\to RL}$ from the signal ℓ to the receptor RL and the static gain from the latter to the output x^*, $\bar{g}_{RL\to x}$: $\bar{g}_{\ell\to x} = \bar{g}_{\ell\to RL}\bar{g}_{RL\to x}$. The static gain $\bar{g}_{RL\to x} = X_{\mathrm{T}} f(1-f)/(R_{\mathrm{T}} p)$ increases with the number of readout molecules X_{T}, but decreases with the number of receptor proteins R_{T}. The static gain $\bar{g}_{\ell\to RL} = p(1-p)R_{\mathrm{T}}/\bar{\ell}$ increases with R_{T} because increasing R_{T} amplifies the signal variations. However, increasing R_{T} also decreases $\bar{g}_{RL\to x}$, in such a way that the overall static gain $\bar{g}_{\ell\to x}$ is independent of R_{T}. The dynamic gain depends not only on the static gain, which determines how much a change in a constant input will lead to a change in the output, but also on the dynamics of the input and the response system. In particular, a slow response will dampen the propagation of input fluctuations. Indeed, the gain $\tilde{g}$ increases as the integration time τ_{r} decreases, because a shorter τ_{r} means that the current output is shaped more strongly by the more recent input, which is correlated more strongly with the future input than the input further back into the past. We also note that $\tilde{g}$ decreases with the forecast interval τ.

SNR$_{\text{pred}}$. The inverse of the SNR, SNR$^{-1}_{\text{pred}}$, is given by Eq. 3.51 with the instantaneous correlation coefficient given by Eq. 3.50. The quantity SNR$^{-1}_{\text{pred}}$ is the relative error in predicting the future concentration. The first two terms on the right-hand side of Eq. 3.50 give the error arising from the readout modification noise and the receptor switching noise. In [15], this error is called the 'sampling error' because it is the sensing error that arises because the push-pull network can be viewed and analyzed as a device that samples the ligand-binding state of the receptor in a stochastic and discrete fashion. Importantly, the SNR depends on the noise and the gain squared, see Eq. 3.12. Hence, the sampling error decreases with the number of readout molecules as $1/X_{\text{T}}$ because while the readout-modification noise increases with X_{T}, the gain-squared increases as X_{T}^2. The receptor-switching noise at the level of the receptors, $R_{\text{T}}p(1-p)$, increases with R_{T} (Eq. 3.69). The static gain $\bar{g}_{RL\to x^*}$ from the receptor to x^* decreases as $1/R_{\text{T}}$, but this affects the propagation of both the signal variations, $\tilde{g}^2\sigma^2_\ell \propto \bar{g}^2_{RL\to x}\sigma^2_\ell$ (see Eqs. 3.12 and 3.45), and the receptor-switching noise, $\propto \bar{g}^2_{RL\to x}R_{\text{T}}p(1-p)$ (see Eq. 3.69), from the receptor to the output in the same way; changing R_{T}, therefore, does not change the contribution to the SNR that comes from the propagation of the signal and the noise from the receptor (thus not the ligand, i.e. signal) to the output. However, the overall static gain $\bar{g}_{\ell\to x}$ also depends on the static gain from the ligand to the receptor $\bar{g}_{\ell\to RL}$, which increases with R_{T}, such that SNR$^{-1}_{\text{pred}}$ and hence the sampling error scale with the number of receptor molecules as $1/R_{\text{T}}$. We thus see that increasing X_{T} and R_{T} raises the SNR by increasing the static gain, which helps to lift the signal above the readout-modification and receptor-switching noise. The relative prediction error, Eq. 3.50, depends not only on the sampling error, but also on what is called in [15, 29] the 'dynamical error'. It arises from the dynamics in the input. Even when the number of readout and receptor molecules becomes very large, and the sampling error reduces to zero, there is still a finite error (see Eq. 3.50). This dynamical error is independent of X_{T} and R_{T} and only depends on ratios of timescales. It stems from the fact that readout molecules encode the receptor history and hence the ligand concentration in the past τ_{r}, which will, in general, deviate from the future concentration that the cell aims to predict. This error thus arises from fluctuations in the past input that are not informative about the future input. Indeed, this error is unique to the predictive information I_{pred}; for the past information I_{past}, the full variance $\sigma^2_{x|\eta}$ caused by fluctuations in the past ligand concentration is part of the 'signal' in the signal-to-noise ratio (Eqs. 3.6 and 3.42). The dynamical error only goes to zero when both the receptor-ligand correlation time τ_{c} and the receptor integration time τ_{r} become much shorter than the input correlation time τ_ℓ.

Optimal integration time. The integration time affects the prediction error in three ways. Increasing the integration time τ_{r} enables more time averaging, thus reducing the sampling error. However, increasing the integration time decreases the dynamic gain $\tilde{g}$ (see Eq. 3.45), which makes it harder to lift the signal above the sampling noise. In addition, increasing τ_{r} increases the dynamical error. The interplay between these three effects gives rise to an optimal integration time that minimizes the relative prediction error.

Optimal ratio $X_{\text{T}}/R_{\text{T}}$. Increasing the number of receptor or readout molecules always increases the precision with which the cell can sense the past and predict the future (Eqs. 3.42 and 3.50). However, when the total resource pool is constrained, the cell

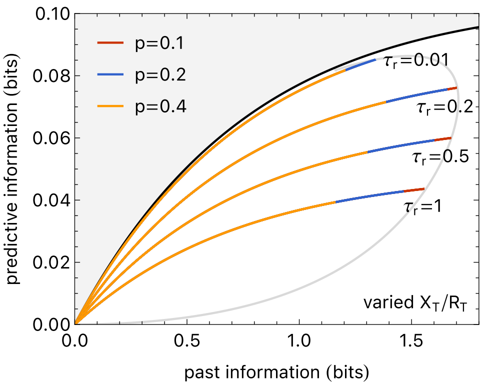


Figure 3.6: The past and predictive information are maximized by the same ratio X_T/R_T and fraction p. The information plane, showing the information bound in black, and the isocost line $C = R_T + X_T = 10^4$ in gray. To construct the colored lines in this figure the ratio X_T/R_T has been varied from zero to a value beyond the optimal value that maximizes I_{past} and I_{pred}. This is done for several values of the receptor occupancy p ($p = 0.1$ in red, $p = 0.2$ in blue, $p = 0.4$ in orange), and for several values of τ_r (indicated in the figure). When X_T/R_T reaches its optimal value, both I_{past} and I_{pred} are maximal. When the ratio is increased further the system moves back to the origin via the same coordinates. Only the integration time τ_r meaningfully distinguishes between strategies that maximize predictive or past information, or that approach the information bound. The reason is that X_T/R_T, together with p, controls the optimal extraction of information that is encoded in the receptor-ligand binding history by ensuring that each receptor is sampled roughly once every receptor correlation time τ_c [13]; the optimal ratio X_T/R_T is thus the same for I_{past} and I_{pred}, but depends on τ_r. The gray isocost line is obtained by varying τ_r, while maximizing for each τ_r the correlation coefficient given by Eq. 3.50; the latter is done by substituting Eq. 3.72 into Eq. 3.50 and numerically optimizing the resulting expression over p. The isocost line gives the region of I_{past} and I_{pred} that is accessible for a given resource cost C. Parameter values are $f = f^{\text{opt}} = 1/2$, $(\sigma_\ell/\bar{\ell})^2 = 10^{-2}$, $\tau_c/\tau_\ell = 10^{-2}$.

has to choose whether it makes more receptors or more readout molecules. To find the optimal ratio of read-out to receptor molecules we write the cost function (Eq. 3.52) in generic form, $C = AR_T + BX_T$, and express X_T and R_T in terms of the total cost C and the ratio X_T/R_T:

$$X_T = C\frac{X_T/R_T}{A + BX_T/R_T}, \tag{3.70}$$

$$R_T = C\frac{1}{A + BX_T/R_T}. \tag{3.71}$$

The factors A and B set the cost of receptors and readout molecules, respectively. Substituting these expressions for X_T and R_T into Eq. 3.50, setting the derivative of the resulting expression with respect to $X_\mathrm{T}/R_\mathrm{T}$ to zero, and solving for $X_\mathrm{T}/R_\mathrm{T}$ gives

$$\begin{aligned}(X_\mathrm{T}/R_\mathrm{T})^\mathrm{opt} &= \sqrt{\left(1+\frac{\tau_\mathrm{r}}{\tau_\mathrm{c}}\right)\frac{p}{1-p}\frac{1}{f(1-f)}\frac{A}{B}},\\ &= 2\sqrt{p/(1-p)}\sqrt{1+\tau_\mathrm{r}/\tau_\mathrm{c}},\end{aligned} \tag{3.72}$$

where for the second line we used $A = B = 1$ and $f = f^\mathrm{opt} = 1/2$. This is the optimal ratio of readout to receptor molecules in the push-pull network, given an integration time τ_r and a steady state fraction of ligand-bound receptors p. This optimal ratio $(X_\mathrm{T}/R_\mathrm{T})^\mathrm{opt}$ maximizes, for a given τ_r and p, not only the predictive information, but also the past information. The reason is that the ratio $X_\mathrm{T}/R_\mathrm{T}$ determines, together with τ_r and p, the interval Δ for sampling the ligand-binding state of the receptor: when the ratio $X_\mathrm{T}/R_\mathrm{T}$ obeys Eq. 3.72, the readout molecules sample each receptor molecule roughly once every correlation time: $\Delta \sim \tau_\mathrm{c}$ [13, 15]. Equation 3.72 is thus a statement about optimally extracting the information that is encoded in the receptor-ligand binding history, concerning both the past information and the predictive information. This can be seen from Eqs. 3.42 and 3.50 and is further illustrated in Fig. 3.6.

4

THE *E. coli* CHEMOTAXIS NETWORK PREDICTS OPTIMALLY IN SHALLOW GRADIENTS

Organisms that navigate their environment with directional persistence, such as Escherichia coli during chemotaxis, experience non-Markovian signal statistics. In this chapter, we show that for typical signals encountered by E. coli during chemotaxis, the optimal integration kernel is a perfectly adaptive derivative-taking kernel, precisely as E. coli employs. Therefore, its chemotaxis network can in principle reach the information bound. However, we again find that reaching the bound is exceedingly costly, because a limited resource availability prevents the system from taking an instantaneous derivative. As we found for the push-pull network in the previous chapter, the most recent information is the most predictive but also the most costly. Computing the past and predictive information for the E. coli chemotaxis network directly from experimental data shows that this system is indeed far from the information bound. Yet, it predicts concentration changes optimally in shallow gradients under a resource constraint.

The contents of this chapter have been published in Proceedings of the National Academy of Sciences **120**, e2303078120 (2023) [38].

In the previous chapter, we have shown that the push-pull network can optimally predict Markovian signals. But not all signals are expected to be Markovian. Especially organisms that navigate through an environment with directional persistence will sense a non-Markovian signal, as generated by their own motion. Moreover, when these organisms need to climb a concentration gradient, as *E. coli* during chemotaxis, then knowing the change in the concentration is arguably more useful than knowing the concentration itself. Indeed, it is well known that the kernel of the *E. coli* chemotaxis system detects the (relative) change in the ligand concentration by taking a temporal derivative of the concentration [19]. However, as our analysis in Chapter 2 has revealed, the converse statement is more subtle. If the system needs to predict the (future) change in the signal, v_τ, then the optimal kernel is not necessarily one that is based on the derivative only: in general, the optimal kernel uses a combination of the signal value and its derivative. This raises the question: can the *E. coli* chemotaxis network reach the information bound for the signals it encounters during chemotaxis?

The *E. coli* chemotaxis system can respond to concentrations that vary between the dissociation constants of the inactive and active state of the receptors, which differ by several orders of magnitude [52]. This range of possible background concentrations is much larger than the typical concentration change over the orientational correlation time of the bacterium. In this chapter, we show that in this regime the optimal kernel is a perfectly adaptive, derivative-taking kernel that is insensitive to the current signal value, precisely like that of the *E. coli* chemotaxis system [19, 53–55]. Our analysis thus predicts that this system has an adaptive kernel because this is the optimal kernel for predicting concentration derivatives over a broad range of background concentrations. Moreover, this observation suggests that *E. coli* may reach the information bound. However, we again we ask ourselves the question whether this network can reach the information bound when considering physical resource costs such as proteins, energy, and time?

In what follows, we first derive the biologically relevant parameter regime of the input signal for *E. coli* performing chemotaxis in shallow gradients, based on experimentally measured signal statistics. To study how close the chemotaxis signaling network can come to the information bound for such a signal, we construct a linear model of the chemotaxis network and subsequently optimize its parameters for prediction under a resource constraint. While the network should be able to reach the bound, we find that reaching the information bound would be prohibitively costly, because it requires taking an instantaneous derivative. The optimal system that maximizes the predictive information under a resource constraint is therefore not at the information bound, emerging from a trade-off between taking a derivative that is recent and one that is reliable. In the following section we study the effect of the chemical power cost on the ability of the chemotaxis network to reach the information bound. Finally, computing the past and predictive information directly from experimental data [20] reveals that the *E. coli* chemotaxis system is indeed markedly away from the information bound and that it is optimally designed to predict future concentration changes in shallow gradients. In this regime, it uses its collected past information for prediction most efficiently.

4.1. OPTIMAL PREDICTION IN SHALLOW GRADIENTS

To reveal the signal characteristics that control the shape of the optimal integration kernel for *E. coli* performing chemotaxis in shallow gradients, we again consider the family of stationary Gaussian signals discussed in Section 2.5, generated by [10, 11]

$$\delta\dot{\ell} = v(t), \tag{4.1}$$

$$\dot{v} = -\omega_0^2 \delta\ell(t) - v(t)/\tau_v + \eta_v(t), \tag{4.2}$$

where $\delta\ell$ is the deviation of the ligand concentration from its mean $\bar{\ell}$, v its derivative, τ_v a relaxation time, and η_v is a Gaussian white noise process which drives the signal dynamics; $\omega_0^2 = \sigma_v^2/\sigma_\ell^2$ sets the ratio of the variance in the derivative of the concentration, σ_v^2, to that of its value σ_ℓ^2. This model covers a range of signal behaviors, allowing us to elucidate the signal characteristics that control the optimal shape of the integration kernel. We first shortly summarize the main findings of Section 2.5, before we show how this model can describe the biologically relevant regime of chemotaxis in shallow gradients.

In Section 2.5 we have shown that the optimal encoding that enables the system to reach the information bound on predicting the future concentration change v_τ, is based on a linear combination of the current concentration $\ell(t)$ and its derivative $v(t)$ [10, 11]:

$$x(t) = a\frac{\delta\ell(t)}{\sigma_\ell} + b\frac{v(t)}{\sigma_v} + \eta_x(t). \tag{4.3}$$

This can be understood by noting that while the signal of Eqs. 4.1 and 4.2 is non-Markovian in the space of ℓ, the concentration *E. coli* actually senses, it is Markovian in ℓ and v: all the information on the future signal is contained in the current concentration and its derivative. To reach the information bound, the coefficients must obey

$$a^{\text{opt}} = G\frac{\langle\delta\ell(0)\delta v(\tau)\rangle}{\sigma_\ell\sigma_v} \equiv G\rho_{\ell_0 v_\tau}, \tag{4.4}$$

$$b^{\text{opt}} = G\frac{\langle\delta v(0)\delta v(\tau)\rangle}{\sigma_v^2} \equiv G\rho_{v_0 v_\tau}. \tag{4.5}$$

Here, G is the gain, which together with the noise $\sigma_{\eta_x}^2$ sets the scale of I_{pred} and I_{past}, $\rho_{\ell_0 v_\tau}$ is the cross-correlation coefficient between the current concentration value ℓ_0 and the future concentration derivative v_τ and $\rho_{v_0 v_\tau}$ that between the current and future derivative (Section 2.5). These expressions can be understood intuitively: if the future signal derivative that needs to be predicted is correlated with the current signal derivative, it is useful to include in the prediction strategy the latter, leading to a non-zero value of b^{opt}. Perhaps more surprisingly, if the future signal derivative is also correlated with the current signal *value*, then the system can enhance the prediction accuracy by also including the current signal value, yielding a non-zero a^{opt}. Clearly, in general, to optimally predict the future signal change, the system should base its prediction on both the current signal value and its derivative.

The degree to which the optimal system bases its prediction on the current signal value versus the current derivative depends on the relative magnitudes of $\rho_{\ell_0 v_\tau}$ and $\rho_{v_0 v_\tau}$, respectively (Eqs. 4.4 and 4.5). These correlation coefficients are set by the spatial structure of the environment in combination with the swimming behavior of the cell. We assume that the spatial structure of the environment is a shallow exponential concentration gradient, where the concentration follows $\ell(x,t) = \bar{\ell}e^{gx(t)}$, with gradient steepness g and position along the gradient direction $x(t)$. The statistics of *E. coli*'s swimming behavior in such an environment have been measured experimentally [20], yielding an exponentially decaying correlation of the cell's velocity in the $x-$direction,

$$\langle \delta v_x(0)\delta v_x(t)\rangle \simeq \sigma^2_{v_x} e^{-\lambda t}. \tag{4.6}$$

Here, correlations decay over the timescale λ^{-1} corresponding to the orientational correlation time of the cell, and $\sigma^2_{v_x}$ is the variance of the cell's velocity. The shallow gradient assumption entails that on average, the cell's velocity is not meaningfully affected by the gradient ($\langle v_x(t)\rangle \approx 0$), such that the cell performs an unbiased random walk over long timescales.

To map the cell's velocity onto the change in concentration $v(t)$ that it encounters, we consider again the spatial structure of the environment $\ell(x,t) = \bar{\ell}e^{gx(t)}$, and use that $v(t) \equiv \frac{d\ell}{dt} = \frac{d\ell}{dx}\frac{dx}{dt}$, where $\frac{dx}{dt} = v_x(t)$. We then obtain $v(t) = g\ell(t)v_x(t)$, which for small concentration deviations $\delta\ell = \ell(t) - \bar{\ell}$ can be approximated to first order as $v(t) \approx g\bar{\ell}v_x(t)$, since $\langle v_x(t)\rangle \approx 0$. In shallow gradients, the encountered change in concentration therefore also decays exponentially,

$$\langle \delta v(0)\delta v(t)\rangle = g^2\bar{\ell}^2 \langle \delta v_x(0)\delta v_x(\tau)\rangle = g^2\bar{\ell}^2\sigma^2_{v_x}e^{-\lambda t}. \tag{4.7}$$

The dynamics corresponding to this correlation function are those generated by a simple Ornstein-Uhlenbeck process in the concentration change:

$$\delta\dot{\ell} = v(t), \tag{4.8}$$

$$\dot{v} = -\lambda v(t) + g\bar{\ell}\sigma_{v_x}\sqrt{2\lambda}\xi(t), \tag{4.9}$$

where $\xi(t)$ is a unit white noise process modeling the stochasticity in the cell's swimming behavior.

To now learn how *E. coli* should optimally predict the signals it encounters in shallow concentration gradients, with the dynamics given in Eqs. 4.8 and 4.9, we map these dynamics onto the more general non-Markovian signal model discussed above and in Chapter 2 (Eqs. 4.1 and 4.2). Mathematically, it is clear that the latter maps onto the former when we set the correlation time $\tau_v = \lambda^{-1}$, the noise strength $\langle \eta^2_v\rangle = 2\lambda g^2\bar{\ell}^2\sigma^2_{v_x}$, and we take the limit $\omega_0 \to 0$. But what does this mean biologically? Firstly, the swimming behavior of the cell completely defines the correlation time of the encountered concentration changes. Secondly, the strength of the fluctuations is set by the swimming statistics as well as the background concentration and gradient steepness, such that the variance in the concentration changes is given by $\sigma^2_v = g^2\bar{\ell}^2\sigma^2_{v_x}$. Finally, and perhaps most interestingly, the parameter ω_0 bounds the total concentration range of the general non-Markovian signal, since we have $\sigma_\ell = \sigma_v/\omega_0$ (see also Eq. 2.57). The fact that this parameter vanishes thus means that the concentration change over the timescale τ_v is much smaller than the range of possible concentrations σ_ℓ that the bacterium can experience, i.e. $\sigma_v\tau_v \ll \sigma_\ell$ and hence $\omega_0 \ll \tau_v^{-1}$. Interestingly, it is known that *E. coli* can sense over a wide range of background concentrations set by the dissociation constants of the inactive and active receptors, which for the Tar-MeAsp receptor ligand combination respectively are $K_D^I = 18\mu\text{M}$ and $K_D^A = 2900\mu\text{M}$ [56–58]. This indicates that *E. coli* operates in a regime where the typical concentration change over a run is much smaller than the range of background concentrations over which it can sense, corresponding to the regime considered here, with $\omega_0 \ll \tau_v^{-1}$. We also note here that while in this limit the statistics of v becomes Markovian, that of the concentration ℓ remains non-Markovian.

Finally, let us determine the required correlation coefficients (Eqs. 4.4 and 4.5) as a function of the prediction interval τ for the signal defined by Eqs. 4.1 and 4.2 in the limit $\omega_0 \to 0$. Using Eq. 2.63

$$\lim_{\omega_0\to 0}\begin{pmatrix} \frac{\langle\delta\ell(\tau)\delta\ell(0)\rangle}{\sigma_\ell^2} & \frac{\langle\delta\ell(\tau)\delta v(0)\rangle}{\sigma_\ell\sigma_v} \\ \frac{\langle\delta v(\tau)\delta\ell(0)\rangle}{\sigma_\ell\sigma_v} & \frac{\langle\delta v(\tau)\delta v(0)\rangle}{\sigma_v^2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & e^{-\tau/\tau_v} \end{pmatrix}. \qquad (4.10)$$

We thus find that, in this regime where the range of concentrations is very large relative to the concentration change over the course of a run ($\sigma_\ell \gg \sigma_v\tau_v$), the current concentration is no longer correlated with the future derivative, i.e. the cross-correlation coefficient vanishes $\rho_{\ell_0 v_\tau} \to 0$. In this case, the weight given to the current concentration a^{opt} becomes zero (Eqs. 4.3 and 4.4): the optimal kernel has become a perfectly adaptive, derivative-taking kernel, as observed for the *E. coli* chemotaxis network. Our analysis thus suggests that *E. coli* has an adaptive kernel precisely because this is the optimal kernel for predicting the future derivative in the navigation regime of *E. coli*, corresponding to $\omega_0 \ll \tau_v^{-1}$.

We emphasize that while we have derived the results above for the class of signals defined by Eqs. 4.1 and 4.2, the idea is far more generic. In particular, while we do not know the temporal structure of the ligand statistics that *E. coli* experiences in general, we do know it can detect concentration changes over a range of background concentrations that is much larger than the typical concentration change over a run [52]. In this regime, the correlation between the concentration value and its future change is likely to be very small. As our analysis shows, a perfectly adaptive kernel then emerges naturally from the requirement to predict the future concentration change.

In summary, our model, as defined by Eqs. 4.1 and 4.2, yields signal dynamics as encountered by *E. coli* in the regime of shallow gradients in the limit that $\omega_0 \ll \tau_v^{-1}$. This corresponds to a regime where the signal change over a run is much smaller than the range of background concentrations, which, as we argue, is the regime of *E. coli*. In this regime, the optimal kernel is a perfectly adaptive kernel, as that of the *E. coli* chemotaxis network.

4.2. THE *E. coli* CHEMOTAXIS NETWORK

The above analysis indicates that the chemotaxis system is ideally designed to predict the future concentration change, because its integration kernel is nearly perfectly adaptive [19, 53–55, 59]. But how close can this system come to the information bound for the signals specified by Eqs. 4.8 and 4.9?

An increasing body of evidence suggests that in the *E. coli* chemotaxis system, re-

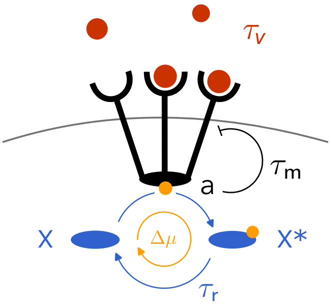


Figure 4.1: The *E. coli* chemotaxis network. Derivative-taking networks, like the *E. coli* chemotaxis system, can optimally predict the future derivative of non-Markovian signals, with correlation time τ_v (Eqs. 4.6-4.9). The chemotaxis system is similar to a push-pull network (Fig. 3.1), yet augmented with receptor cooperativity and with negative feedback on the receptor activity via methylation on a timescale τ_m.

ceptors cooperatively control the activity of the kinase CheA (Fig. 4.1) [60–63]. Furthermore, the kinase activity is adaptive due to the methylation of inactive receptors [19, 64]. A widely used approach to describe the effects of receptor cooperativity and methylation on kinase activity, has been to employ the Monod-Wyman-Changeux (MWC) model [20, 50, 52, 57, 62, 65–67]. We will follow this approach and, more specifically, model the chemotaxis system as described by Tu and colleagues [58]. In this model, each receptor can switch between an active and inactive conformational state. Moreover, receptors are partitioned into clusters of equal size N. In the spirit of the MWC model, receptors within a cluster switch conformation in concert, so that each cluster is either active or inactive [65]. Furthermore, it is assumed that receptor-ligand binding and conformational switching are faster than the other timescales in the system (for an alternative model including ligand binding dynamics see Appendix 4.B). The probability for the kinase, i.e. the receptor cluster, to be active, is then described by:

$$a(\ell, m) = \frac{1}{1 + \exp(\Delta F_T(\ell, m))}, \tag{4.11}$$

where $\Delta F_T(\ell, m)$ is the total free-energy difference between the active and inactive state, which is a function of the ligand concentration $\ell(t)$ and the methylation level of the cluster $m(t)$. The simplest model adopted here assumes a linear dependence of the total free-energy difference on the free-energy difference arising from ligand binding and methylation:

$$\Delta F_T(\ell, m) = -\Delta E_0 + N(\Delta F_\ell(\ell) + \Delta F_m(m)), \tag{4.12}$$

where the free-energy difference due to ligand binding is

$$\Delta F_\ell(\ell) = \ln(1 + \ell(t)/K_\mathrm{D}^\mathrm{I}) - \ln(1 + \ell(t)/K_\mathrm{D}^\mathrm{A}). \tag{4.13}$$

Between the two states the cluster has an altered dissociation constant, which is denoted K_D^I for the inactive state, and K_D^A for the active state. The free-energy difference due to methylation has been experimentally shown to depend approximately linearly on the methylation level [62]:

$$\Delta F_m(m) = \tilde{\alpha}(\bar{m} - m(t)). \tag{4.14}$$

We assume that inactive receptors are irreversibly methylated, and active receptors irreversibly demethylated, with zero-order ultrasensitive kinetics [36, 58, 68]. The dynamics of the methylation level of the i^{th} receptor cluster is then given by:

$$\dot{m}_i = (1 - a_i(\ell, m_i))k_R - a_i(\ell, m_i)k_B + B_{m_i}(a_i)\xi(t), \tag{4.15}$$

with $B_m^{(i)}(a_i) = \sqrt{(1 - a_i(\ell, m_i))k_R + a_i(\ell, m_i)k_B}$, and unit white noise $\xi(t)$. These dynamics indeed give rise to perfect adaptation, since from this equation we find that the steady state cluster activity is given by $p \equiv \bar{a} = 1/(1 + k_B/k_R)$, independent of the ligand concentration.

Finally, active receptors catalyze phosphorylation of read-out molecules, and phosphorylated read-out molecules decay at a constant rate. We have

$$\dot{x}^* = \sum_{i=1}^{R_{\text{T}}} a_i(t)(X_{\text{T}} - x^*(t))k_{\text{f}} - x^*(t)k_{\text{r}} + B_x(a_i, x^*)\xi(t), \tag{4.16}$$

where R_{T} is the total number of receptor *clusters*. The steady state fraction of phosphorylated read-outs is given by $f \equiv \bar{x}^*/X_{\text{T}} = (1 + k_{\text{r}}/(k_{\text{f}}R_{\text{T}}p))^{-1}$.

LINEAR DYNAMICS

We again perform a linear noise approximation (see Appendix 3.B) and define all variables in terms of deviations from their mean: $\delta\ell(t) = \ell(t) - \bar{\ell}$, $\delta m(t) = m(t) - \bar{m}$ and $\delta a(t) = a(t) - p$. The linear form of this chemotaxis model has previously been studied in for example [36] and [58]. We obtain for the linear dynamics of the i^{th} cluster activity

$$\delta a_i(t) = \alpha\delta m_i(t) - \beta\delta\ell(t), \tag{4.17}$$

where $\alpha = \tilde{\alpha}Np(1-p)$ and $\beta = \kappa Np(1-p)$, with $\kappa = (\bar{\ell} + K_{\text{D}}^{\text{I}})^{-1} - (\bar{\ell} + K_{\text{D}}^{\text{A}})^{-1}$. For the methylation on the i^{th} cluster and for the readout dynamics we then obtain, as a function of $\delta a(t)$,

$$\delta\dot{m}_i = -\delta a_i(t)/(\alpha\tau_{\text{m}}) + \eta_{m_i}(t), \tag{4.18}$$

$$\delta\dot{x}^* = \gamma\sum_{i=1}^{R_{\text{T}}} \delta a_i(t) - \delta x^*(t)/\tau_r + \eta_x(t), \tag{4.19}$$

where we have introduced the relaxation times $\tau_{\text{m}} = (\alpha(k_R + k_B))^{-1}$ for methylation and $\tau_{\text{r}} = (R_{\text{T}}pk_{\text{f}} + k_{\text{r}})^{-1}$ for phosphorylation. We have further defined the rate at which an active cluster phosphorylates the readout CheY: $\gamma = X_{\text{T}}f(1-f)/(pR_{\text{T}}\tau_{\text{r}})$. Substituting the expression for δa_i (Eq. 4.17) into Eqs. 4.18 and 4.19, and expressing the dynamics in terms of the methylation on all clusters gives

$$\frac{d}{dt}\left(\sum_{i=1}^{R_\mathrm{T}} \delta m_i\right) = -\sum_{i=1}^{R_\mathrm{T}} \delta m_i/\tau_\mathrm{m} + q\delta\ell(t)/(\alpha\tau_\mathrm{m}) + \eta_m(t), \tag{4.20}$$

$$\delta\dot{x}^* = -\delta x^*(t)/\tau_\mathrm{r} - \gamma q \delta\ell(t) + \gamma\alpha \sum_{i=1}^{R_\mathrm{T}} \delta m_i(t) + \eta_x(t), \tag{4.21}$$

with $q = R_\mathrm{T}\beta$ (see Eq. 4.17 for β). The rescaled white noise η_m is the sum of the methylation noise on all receptor clusters, where we have assumed that the methylation noise is independent between clusters. The noise strengths of the methylation and phosphorylation noise respectively are

$$\langle\eta_m^2\rangle = 2R_\mathrm{T}p(1-p)/(\alpha\tau_\mathrm{m}), \tag{4.22}$$

$$\langle\eta_x^2\rangle = 2X_\mathrm{T}f(1-f)/\tau_\mathrm{r}. \tag{4.23}$$

Parameter values

A large body of work has studied the parameters of the MWC model for the *E. coli* chemotaxis system. We have listed the parameters relevant for our model in table 4.1. We choose the background concentration $\bar{\ell}$ to be in between K_D^I and K_D^A, at $\bar{\ell} = 100\mu$M.

Table 4.1: Measured *E. coli* chemotaxis parameter values.

Parameter	*Value*	*Source*	*Description*
K_D^I	18μM	[56–58]	MeAsp-Tar inactive dissociation constant
K_D^A	2900μM	[56–58]	MeAsp-Tar active dissociation constant
N	~ 12	[20, 35]	Inferred cluster size (see Section 4.5)
$\tilde{\alpha}$	$2k_\mathrm{B}T$	[62]	Free energy change per methyl group
p	0.3	[20, 57, 62]	Steady state activity at 22°C
τ_r	$\sim 0.1s$	[13, 20, 56, 69]	Phosphorylation timescale

In this work we analyze the impact of the methylation timescale τ_m, and the number of receptor clusters and readout molecules R_T and X_T, on the past and predictive information. We therefore do not set them to a fixed value, but experimental estimates are listed in table 4.2.

Table 4.2: Approximate *E. coli* chemotaxis timescales and abundances.

Parameter	*Value*	*Source*	*Description*
τ_m	$\sim 10s$	[19, 20, 62]	Adaptation time
Tsr+Tar	14000, 3300	[49]	Rich medium; RP437, OW1 strain
Tsr+Tar	24000, 37000	[49]	Minimal medium; RP437, OW1 strain
CheY	8200, 1400	[49]	Rich medium; RP437, OW1 strain
CheY	6300, 14000	[49]	Minimal medium; RP437, OW1 strain

MODEL STATISTICS

Here we we derive the variance in the network output, the signal to noise ratios, and the correlation coefficient between current output and the future signal for the chemotaxis network. As in Chapter 3, we exploit the spectral properties of the network to derive these quantities, following the procedure outlined in Appendix 3.C. We consider the system to sense the non-Markovian ligand concentration defined by Eqs. 4.1 and 4.2. Such a signal is characterized by both its concentration and derivative. The power spectrum of the concentration is given by

$$S_\ell(\omega) = \frac{2\sigma_v^2/\tau_v}{(\omega^2 + ((2\tau_v)^{-1} + \rho)^2)(\omega^2 + ((2\tau_v)^{-1} - \rho)^2)}, \tag{4.24}$$

where $\rho = \sqrt{(4\tau_v^2)^{-1} - \omega_0^2}$. The power spectrum of the derivative is related to that of the concentration as $S_v(\omega) = \omega^2 S_\ell(\omega)$, and the cross-spectra are $S_{\ell\to v}(\omega) = S_{v\to\ell}(-\omega) = i\omega S_\ell(\omega)$.

The linear dynamics of the chemotaxis signaling network are fully determined by the following matrices (also see Appendix 3.B)

$$\boldsymbol{G} = q\begin{pmatrix} 1/(\alpha\tau_\mathrm{m}) & 0 \\ -\gamma & 0 \end{pmatrix}, \tag{4.25}$$

$$\boldsymbol{J} = \begin{pmatrix} -1/\tau_\mathrm{m} & 0 \\ \alpha\gamma & -1/\tau_\mathrm{r} \end{pmatrix}, \tag{4.26}$$

$$\boldsymbol{B} = \begin{pmatrix} \sqrt{2R_\mathrm{T}p(1-p)/(\alpha\tau_\mathrm{m})} & 0 \\ 0 & \sqrt{2X_\mathrm{T}f(1-f)/\tau_\mathrm{r}} \end{pmatrix}, \tag{4.27}$$

where we again have signal gain matrix $\boldsymbol{G}$ describing the strength by which the signal impacts each species, the Jacobian $\boldsymbol{J}$ of the signaling network, and the matrix $\mathcal{B}$ of noise strengths.

The Fourier transform of the matrix exponential of the Jacobian is (also see Appendix 3.C)

$$\begin{aligned} \mathcal{F}\{e^{\boldsymbol{J}t}\} &= (i\omega\mathbb{I}_n - \boldsymbol{J})^{-1} \\ &= \begin{pmatrix} \frac{1}{1/\tau_\mathrm{m}+i\omega} & 0 \\ \frac{\alpha\gamma}{(1/\tau_\mathrm{m}+i\omega)(1/\tau_\mathrm{r}+i\omega)} & \frac{1}{1/\tau_\mathrm{r}+i\omega} \end{pmatrix}, \end{aligned} \tag{4.28}$$

which allows us to determine the gain matrix via $\mathbb{K}(\omega) = \mathcal{F}\{e^{\boldsymbol{J}t}\}\boldsymbol{G}$, and the noise matrix using $\mathbb{N}(\omega) = \mathcal{F}\{e^{\boldsymbol{J}t}\}\boldsymbol{B}$ (see Eq. 3.65 and Eq. 3.66).

To gain more insight in the way in which the network maps the signal onto its output, we first study the integration kernels of the system. The integration kernel from ligand concentration to output is given by the inverse Fourier transform of element $(2,1)$ of the gain matrix $\mathbb{K}(\omega)$, which is

$$k(t) \equiv \mathcal{F}^{-1}\{K_{\ell\to x}(\omega)\} = \kappa N f(1-f)(1-p)X_{\mathrm{T}}\frac{1}{1-\tau_{\mathrm{r}}/\tau_{\mathrm{m}}}\left(\frac{1}{\tau_{\mathrm{m}}}e^{-t/\tau_{\mathrm{m}}} - \frac{1}{\tau_{\mathrm{r}}}e^{-t/\tau_{\mathrm{r}}}\right), \quad (4.29)$$

with $\kappa = (\bar{\ell} + K_{\mathrm{D}}^{\mathrm{I}})^{-1} - (\bar{\ell} + K_{\mathrm{D}}^{\mathrm{A}})^{-1}$. Due to the adaptive nature of the network, the static gain from ligand concentration to output is zero: $\bar{g}_{\ell\to x} = \int_0^\infty k(t)dt = 0$; the long-time response to a step change of the input concentration is zero. The kernel does indeed not change the output based on the input concentration directly, but instead takes a (time-averaged) derivative of the input (Fig. 4.2A). It is therefore useful to consider the kernel that maps the signal derivative onto the output. This kernel can be found by rearranging the expression for the output of a linear signaling network, Eq. 4.19. Disregarding the noise terms and integrating by parts gives

$$\int_{-\infty}^{0} k(-t)\ell(t)dt = K(-t)\ell(t)|_{-\infty}^{0} - \int_{-\infty}^{0} K(-t)v(t)dt, \quad (4.30)$$

where $v(t) \equiv \dot{\ell}$ and $K(t)$ is the primitive of $k(t)$. To make progress we first determine $K(t)$,

$$K(t) = \kappa N f(1-f)(1-p)X_{\mathrm{T}}\frac{1}{1-\tau_{\mathrm{r}}/\tau_{\mathrm{m}}}\left(-e^{-t/\tau_{\mathrm{m}}} + e^{-\tau/\tau_{\mathrm{r}}}\right). \quad (4.31)$$

The form of $K(t)$ is that of a simple exponential kernel with a delay (Fig. 4.2B). We thus have both $K(0) = 0$ and $K(\infty) = 0$. It is now clear that the convolution over the ligand concentration simply maps onto the convolution over its derivative as

$$\int_{-\infty}^{0} k(-t)\ell(t)dt = -\int_{-\infty}^{0} K(-t)v(t)dt. \quad (4.32)$$

The static gain of $K(t)$ is

$$\bar{g}_{v\to x} = \int_0^\infty K(t)dt = q\gamma\tau_{\mathrm{r}}\tau_{\mathrm{m}} = \tau_{\mathrm{m}}\kappa N X_{\mathrm{T}} f(1-f)(1-p). \quad (4.33)$$

The gain thus increases with the number of receptors per cluster, N, the number of readout molecules, X_{T}, and notably, with the adaptation time τ_{m}. This static gain from signal derivative to network output is a useful quantity which we will use to describe the other statistics of the network below.

To compute the past and predictive information, we need to determine the variance in the output, the SNR, and the correlation between the current output and the future ligand derivative. To this end we require the power spectrum of the output, and the cross-spectrum from output to future derivative. For the power spectrum of the output we use Eq. 3.66 with Eqs. 4.25-4.28 to find

$$S_x(\omega) = \frac{q^2\gamma^2\omega^2}{(\tau_{\mathrm{r}}^{-2}+\omega^2)(\tau_{\mathrm{m}}^{-2}+\omega^2)}S_\ell(\omega) + \frac{\alpha^2\gamma^2\langle\eta_m^2\rangle}{(\tau_{\mathrm{r}}^{-2}+\omega^2)(\tau_{\mathrm{m}}^{-2}+\omega^2)} + \frac{\langle\eta_x^2\rangle}{\tau_{\mathrm{r}}^{-2}+\omega^2}. \quad (4.34)$$

From this power spectrum we can see that the network is a band-pass filter, where the gain is maximal in the frequency range $\tau_{\mathrm{m}}^{-1} < \omega < \tau_{\mathrm{r}}^{-1}$. Both for $\omega \gg \tau_{\mathrm{r}}^{-1}$ and $\omega \ll \tau_{\mathrm{m}}^{-1}$ the

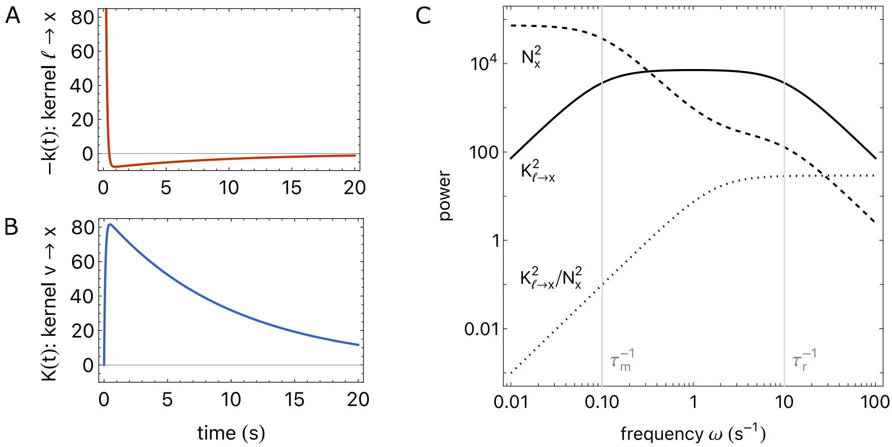


Figure 4.2: Integration kernel and power spectra. (A) The integration kernel $k(t)$ takes a temporal derivative by weighing the most recent signal values with an opposite sign from the preceding ones. (B) The integration kernel $K(t)$ from the derivative of the input concentration to the network output. The kernel $K(t)$ is the primitive of $k(t)$, and its static gain is proportional to the adaptation timescale τ_{m}. (C) Frequency dependent gain $K^2_{\ell\to x}(\omega)$, frequency dependent noise $N_x^2(\omega)$, and their ratio, as a function of frequency. The chemotaxis network is a band-pass filter, the frequencies that are passed through are set by τ_{r} on the high end and τ_{m} on the low end. At low frequencies, the methylation noise dominates. The number of receptor clusters is set equal to the number of readout molecules $R_{\mathrm{T}} = X_{\mathrm{T}} = 5000$. Parameters are $\tau_{\mathrm{r}} = 0.1$s, $\tau_{\mathrm{m}} = 10$s, $\tilde{a} = 2$, $N = 12$, $K_{\mathrm{D}}^{\mathrm{I}} = 18\mu$M, $K_{\mathrm{D}}^{\mathrm{A}} = 2900\mu$M, $\bar{\ell} = 100\mu$M, $f = 0.5$, $p = 0.3$.

gain goes to zero. On long timescales the methylation noise dominates (Fig. 4.2C). The cross-power spectrum between current output and future ligand derivative is given by element (2,2) of the matrix $\mathbb{K}(-\omega)\mathbb{S}_s(\omega)$ which is (also see Eq. 3.67 and Eqs. 4.25-4.28)

$$S_{x\to v}(\omega) = q\gamma \frac{-\omega^2 S_\ell(\omega)}{(\tau_{\mathrm{m}}^{-1} - i\omega)(\tau_{\mathrm{r}}^{-1} - i\omega)}. \tag{4.35}$$

As discussed above, the biologically relevant regime of the input signal is the limit $\sigma_v/\sigma_\ell = \omega_0 \to 0$. We therefore present below the network statistics in this limit. We start by determining the variance in the readout, via the inverse Fourier transform of its power spectrum (Eq. 4.34), i.e.

$$\sigma_x^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_x(\omega)d\omega = \sigma^2_{x|\eta} + \sigma^2_{x|L}, \tag{4.36}$$

with,

$$\begin{aligned}\lim_{\omega_0\to 0}\sigma^2_{x|L} &= X_{\mathrm{T}} f(1-f) + \bar{g}^2_{a\to x}\alpha R_{\mathrm{T}} p(1-p)\frac{1}{1+\tau_{\mathrm{r}}/\tau_{\mathrm{m}}}\\ &= X_{\mathrm{T}} f(1-f)\left(1 + \frac{\tilde{a}NX_{\mathrm{T}} f(1-f)(1-p)^2}{R_{\mathrm{T}}(1+\tau_{\mathrm{r}}/\tau_{\mathrm{m}})}\right),\end{aligned} \tag{4.37}$$

$$\lim_{\omega_0\to 0}\sigma^2_{x|\eta} = \frac{\bar{g}^2_{v\to x}}{(1+\tau_{\mathrm{m}}/\tau_v)(1+\tau_{\mathrm{r}}/\tau_v)}\left(1 + \frac{\tau_{\mathrm{m}}\tau_{\mathrm{r}}}{\tau_v(\tau_{\mathrm{m}}+\tau_{\mathrm{r}})}\right)\sigma_v^2, \tag{4.38}$$

where $\bar{g}_{a\to x} = \gamma\tau_{\mathrm{r}} = X_{\mathrm{T}} f(1-f)/(R_{\mathrm{T}} p)$ is the static gain from receptor activity to readout, and we used the definition of $\alpha = \tilde{\alpha} N p(1-p)$. Because there is no receptor-ligand binding noise, there is also no time averaging as in the push-pull network (and hence no factor depending on $\tau_{\mathrm{r}}/\tau_{\mathrm{c}}$). There is methylation noise on a timescale τ_{m}, but this cannot be time-averaged effectively because the integration time τ_{r} of the push-pull network is shorter than the receptor methylation timescale τ_{m}. The methylation noise can only be averaged out significantly by increasing R_{T}. The contribution from the variance in the signal derivative, σ^2_v, to the output variation σ^2_x, depends on the product of the static gain $\bar{g}^2_{v\to x}$ (Eq. 4.33) and a factor that only depends on ratios of timescales. This term is maximized for $\tau_{\mathrm{r}} \to 0$ and $\tau_{\mathrm{m}} \to \infty$, which is intuitive because in this limit the output is determined by the current input; it subtracts from this current signal the average signal over the past τ_{m}, which for $\tau_{\mathrm{m}} \to \infty$ is however just the baseline signal. In the opposite limit, $\tau_{\mathrm{m}} \to 0$, the gain is indeed zero because the system subtracts from the current signal the same current signal; the system then instantly adapts.

PAST INFORMATION OF THE CHEMOTAXIS NETWORK

The past information is straightforward to compute from the quantities above. The definition of the past information is the same as that for the push-pull network, and is given by Eq. 3.6. The SNR is now given by, using Eqs. 4.37 and 4.38:

$$\begin{aligned}\mathrm{SNR}^{-1}_{\mathrm{past}} &= \frac{\sigma^2_{x|L}}{\sigma^2_{x|\eta}}\\ &= \frac{(1+\tau_{\mathrm{m}}/\tau_v)\,(1+\tau_{\mathrm{r}}/\tau_v)}{\kappa^2 N^2 \tau^2_{\mathrm{m}} \sigma^2_v}\left(1+\frac{\tau_{\mathrm{m}}\tau_{\mathrm{r}}}{\tau_v(\tau_{\mathrm{m}}+\tau_{\mathrm{r}})}\right)^{-1}\left(\frac{1}{X_{\mathrm{T}} f(1-f)(1-p)^2}+\frac{\tilde{\alpha}N}{R_{\mathrm{T}}(1+\tau_{\mathrm{r}}/\tau_{\mathrm{m}})}\right),\end{aligned} \tag{4.39}$$

where $\kappa = (\bar{\ell}+K^{\mathrm{I}}_{\mathrm{D}})^{-1} - (\bar{\ell}+K^{\mathrm{A}}_{\mathrm{D}})^{-1}$. Increasing X_{T} or R_{T} always increases the past information because it reduces the effects of the readout modification and receptor methylation noise. Increasing τ_{m} also monotonically increases the past information because it increases the network gain.

PREDICTIVE INFORMATION OF THE CHEMOTAXIS NETWORK

To compute the predictive information we require the covariance between the current output and the future derivative:

$$\begin{aligned}\lim_{\omega_0\to 0} \langle \delta x(0)\delta v(\tau)\rangle &= \mathcal{F}^{-1}\{S_{x\to v}(\omega)\},\\ &= \frac{-\bar{g}_{v\to x} e^{-\tau/\tau_v}}{(1+\tau_{\mathrm{m}}/\tau_v)(1+\tau_{\mathrm{r}}/\tau_v)}\sigma^2_v = \tilde{g}\sigma^2_v,\end{aligned} \tag{4.40}$$

with the dynamic gain $\tilde{g} \equiv \langle \delta x(0)\delta v(\tau)\rangle / \sigma^2_v$ (see Eq. 3.10) given by

$$\tilde{g} = \frac{-\bar{g}_{v\to x} e^{-\tau/\tau_v}}{(1+\tau_{\mathrm{m}}/\tau_v)(1+\tau_{\mathrm{r}}/\tau_v)}. \tag{4.41}$$

Note that since $\bar{g}_{v\to x} \propto \tau_{\rm m}$ (Eq. 4.33), and generally for the chemotaxis network $\tau_{\rm r} \ll \tau_v$, the dynamic gain depends on the adaptation time as

$$\tilde{g} \sim \tau_{\rm m}/(1+\tau_{\rm m}/\tau_v). \tag{4.42}$$

From Eq. 4.40 it is clear that we can write, just as for the push-pull network (Section 3.4),

$$\langle \delta x(0)\delta v(\tau)\rangle = \langle \delta x(0)\delta v(0)\rangle\, e^{-\tau/\tau_v}. \tag{4.43}$$

This observation, together with Eqs. 3.16 and 3.17, shows that we then only need to specify the instantaneous correlation coefficient $\rho_{xv}(0) \equiv \langle \delta x(0)\delta v(0)\rangle/(\sigma_x\sigma_v)$ to determine the predictive information, as was also the case for the push-pull network. The instantaneous correlation is, using Eq. 4.40 with $\tau = 0$ for the covariance, and using Eqs. 4.36-4.38 for the variance in the output,

$$\begin{aligned}\rho_{xv}^{-2}(0) &= \frac{\sigma^2_{x|L}\sigma^2_v}{\langle \delta x(0)\delta v(0)\rangle^2} + \frac{\sigma^2_{x|\eta}\sigma^2_v}{\langle \delta x(0)\delta v(0)\rangle^2}\\ &= \frac{(1+\tau_{\rm m}/\tau_v)^2(1+\tau_{\rm r}/\tau_v)^2}{\kappa^2 N^2\tau_{\rm m}^2\sigma_v^2}\left(\frac{1}{X_{\rm T} f(1-f)(1-p)^2} + \frac{\tilde{\alpha}N}{R_{\rm T}(1+\tau_{\rm r}/\tau_{\rm m})}\right)\\ &\quad + \left(1+\frac{\tau_{\rm m}}{\tau_v}\right)\left(1+\frac{\tau_{\rm r}}{\tau_v}\right)\left(1+\frac{\tau_{\rm m}\tau_{\rm r}}{\tau_v(\tau_{\rm m}+\tau_{\rm r})}\right),\end{aligned} \tag{4.44}$$

defining the predictive information via Eq. 3.17. This instantaneous correlation coefficient is related to the relative error as (see also Eq. 3.16),

$$\mathrm{SNR}^{-1}_{\rm pred} = \rho_{xv}^{-2}(0)e^{2\tau/\tau_v} - 1. \tag{4.45}$$

In Appendix 4.A we discuss the optimal design that maximizes the signal to noise ratio $\mathrm{SNR}_{\rm pred}$. A more extensive analysis of the signal to noise ratio of the chemotaxis network in terms of the so-called sampling error and dynamical error [13, 15] is left for Chapter 5. In this chapter we focus on the question whether the optimal chemotaxis network under a resource constraint, and the *E. coli* chemotaxis network as observed experimentally, can approach the information bound.

4.3. FINITE RESOURCES PREVENT TAKING AN INSTANTANEOUS DERIVATIVE

To asses whether the optimal system under a resource constraint can reach the information bound, we determine the accessible region of $I_{\rm past}$ and $I_{\rm pred}$ under the resource constraint $C \le NR_{\rm T} + X_{\rm T}$ (see Fig. 4.3) by optimizing over the methylation time $\tau_{\rm m}$ and the ratio of readout over receptor molecules $X_{\rm T}/R_{\rm T}$. We consider the non-Markovian signals specified by Eqs. 4.1 and 4.2 in the physiologically relevant limit $\omega_0 \to 0$, such that the optimal kernel is perfectly adaptive, like that of *E. coli.* We do not optimize over the integration time $\tau_{\rm r}$ because we have assumed that ligand binding is much faster than the other timescales in the system. Therefore, there is no need to time average receptor-ligand binding noise, which means that, in the absence of operating costs, the optimal

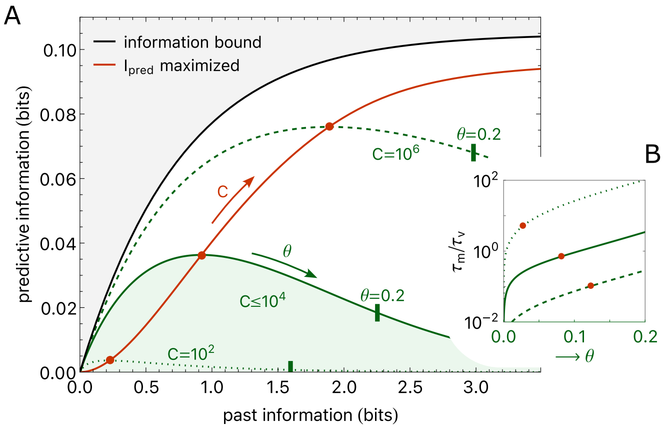


Figure 4.3: Finite resources prevent the chemotaxis system from reaching the information bound. (A) The region of accessible predictive information $I_{\text{pred}} = I(x_0; v_\tau)$ and past information $I_{\text{past}} = I(x; \boldsymbol{L}_p)$ for the chemotaxis system of Eqs. 4.17-4.19 under a resource constraint $C \le NR_{\text{T}} + X_{\text{T}}$, for the signals specified by Eqs. 4.8 and 4.9 (green). The black line shows the information bound at which I_{pred} is maximized for a given I_{past}. The chemotaxis system is not at the information bound, but it does move towards it as C is increased. The red line connects the red points where I_{pred} is maximized for a given resource cost C. The accessible region of I_{pred} and I_{past} under a given resource constraint C is obtained by varying the methylation time τ_{m} and optimizing the ratio of readout over receptor molecules $X_{\text{T}}/R_{\text{T}}$; $g = 4\text{mm}^{-1}$, $N = 12$. (B) The methylation time τ_{m} over the input correlation time τ_v as a function of the distance θ along the three respective isocost lines shown in panel A. The methylation time τ_{m} increases along the isocost line, but there exists an optimal τ_{m} that maximizes the predictive information, marked by the red points; $\theta \to 0$ corresponds to the origin of panel A, $(I_{\text{pred}}, I_{\text{past}}) = (0,0)$; the points where $\theta = 0.2$ along the isocost lines of panel A are marked with a bar. As the resource constraint is relaxed (higher C), the optimal τ_{m} decreases: the system moves towards the information bound, where it takes an instantaneous derivative, corresponding to $\tau_{\text{r}}, \tau_{\text{m}} \to 0$. Forecast interval $\tau = \tau_v$, $\tau_{\text{r}} = 100\text{ms}$ [13, 20, 56]; $\tau_v^{-1} = 0.9\text{s}^{-1}$ and $\sigma_v^2 = g^2 \bar{\ell}^2 \sigma_{v_x}^2$, with $\bar{\ell} = 100\mu\text{M}$ and $\sigma_{v_x}^2 = 157.1\mu\text{m}^2\text{s}^{-2}$ [20]; $f = 0.5$, $p = 0.3$ [20].

integration time $\tau_{\text{r}}^{\text{opt}}$ is zero. In what follows, we set τ_{r} to the value measured experimentally, $\tau_{\text{r}} \approx 100\text{ms}$ [13, 56]. In Appendix 4.B we present a version of the chemotaxis model which does include ligand binding.

Fig. 4.3A shows that the optimal system that maximizes the predictive information I_{pred} under a resource constraint C is markedly away from the information bound. Yet, as the resource constraint is relaxed and C is increased, the optimal system moves towards it. Panel B shows that the methylation time τ_{m} rises along the three respective isocost lines of panel A. It highlights that there exists an optimal methylation time $\tau_{\text{m}}^{\text{opt}}$ that maximizes the predictive information I_{pred}. Moreover, $\tau_{\text{m}}^{\text{opt}}$ decreases as the resource constraint is relaxed. Along the isocost lines $X_{\text{T}}/R_{\text{T}}$ varies only mildly (see Appendix 4.A, Fig. 4.7).

These observations can be understood by noting that the system faces a trade-off between taking a derivative that is recent versus one that is robust. All the information on the future derivative, which the cell aims to predict, is contained in the current derivative of the signal; measuring the current derivative would allow the system to reach the information bound. However, computing the recent derivative is extremely costly. The cell takes the temporal derivative of the ligand concentration at the level of the receptor via two antagonistic reactions that occur on two distinct timescales: ligand binding rapidly deactivates the receptor, while methylation slowly reactivates it [58]. The receptor ligand-occupancy thus encodes the current concentration, the methylation level stores the average concentration over the past τ_{m}, and the receptor activity reflects the difference between the two—the temporal derivative of the signal over the timescale τ_{m}. To obtain an instantaneous derivative, τ_{m} must go to zero. However, the gain $\tilde{g}$, i.e. change in output due to a change in input, scales as $\tilde{g} \sim \tau_{\mathrm{m}}/(1+\tau_{\mathrm{m}}/\tau_{v})$ (Eq. 4.42). Clearly, in the limit $\tau_{\mathrm{m}} \to 0$, the gain becomes zero because the receptor activity instantly adapts to the ligand concentration change. Since the push-pull network downstream of the receptor is a device that samples the receptor stochastically [13, 15], the gain must be raised to lift the signal above this sampling noise. This requires a finite methylation time τ_{m}. The trade-off between a recent derivative and a reliable one gives rise to an optimal methylation time $\tau_{\mathrm{m}}^{\mathrm{opt}}$ that maximizes the predictive information for a given resource cost (see also Appendix 4.A).

The same analysis also explains why the optimal methylation time $\tau_{\mathrm{m}}^{\mathrm{opt}}$ decreases and the predictive information increases when the resource constraint is relaxed (Fig. 4.3). The sampling noise in estimating the average receptor activity decreases as the number of readout molecules increases [13, 15]. A smaller gain is then required to lift the signal above this noise. In addition, a larger number of receptors decreases the noise in the methylation level, which also allows for a smaller gain, and hence a shorter methylation time. These two effects together explain why the optimal methylation decreases with C, scaling as $\tau_{\mathrm{m}}^{\mathrm{opt}} \sim 1/\sqrt{C}$ (Appendix 4.A, Eq. 4.54), and I_{pred} increases with C (Eq. 4.44).

Fig. 4.3A also shows that the past information $I_{\mathrm{past}} = I(x_0; \boldsymbol{L}_p)$ does not return to zero along the contourline of constant resource cost. Along the contourline, the methylation time τ_{m} rises (Fig. 4.3B). While the predictive information I_{pred} exhibits an optimal methylation time $\tau_{\mathrm{m}}^{\mathrm{opt}}$, the past information I_{past} continues to rise with τ_{m} because the system increasingly becomes a copying device, rather than one that takes a temporal derivative.

4.4. Operating costs only slightly alter the optimal chemotaxis network

We have seen that the operating cost moves the push-pull network away from the information bound (Fig. 3.5). To study the effects of operating costs in the chemotaxis network we have to consider the power cost of both phosphorylation and methylation. We therefore extend the cost function to include the cost of methylation,

$$C = \lambda(NR_{\mathrm{T}} + X_{\mathrm{T}}) + c_1 X_{\mathrm{T}} \Delta\mu_{\mathrm{ATP}}/\tau_{\mathrm{r}} + c_2 R_{\mathrm{T}} \Delta\mu_{\mathrm{SAM}}/\tau_{\mathrm{m}}, \tag{4.46}$$

where$\Delta\mu_{\mathrm{ATP}}$ and $\Delta\mu_{\mathrm{SAM}}$ are the free energy costs associated with ATP and S-adenosyl methionine (SAM) hydrolysis, respectively. The methylation flux is set by $R_{\mathrm{T}}/\tau_{\mathrm{m}}$, c_2 is the relative cost of synthesizing components versus the power cost of methylation. To

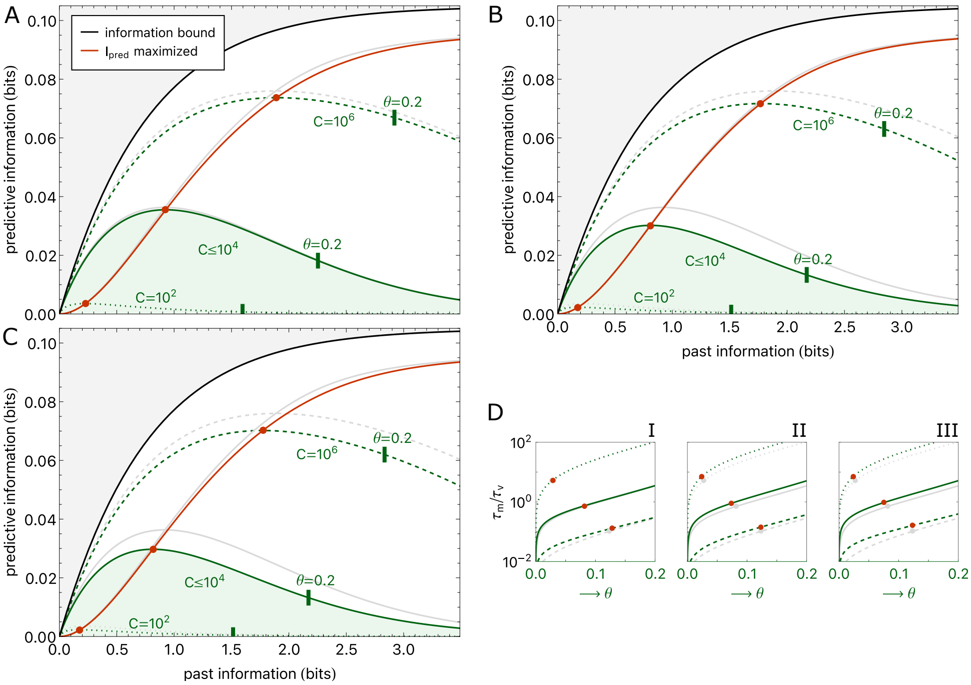


Figure 4.4: Including methylation cost affects the accessible regions and optimal strategies only marginally. (A) Only the methylation cost is included in C ($c_1 = 0$, $c_2 > 0$ in Eq. 4.46). In green, the region of accessible predictive and past information in the chemotaxis network under the resource constraint. The black line is the information bound; the red dots mark the points where I_{pred} is maximized under a resource constraint C; the red line connects these points for increasing C. The gray lines indicate the case without methylation cost, corresponding to Fig. 4.3A. The forecast interval $\tau = \tau_v$. The estimate of c_2 is based on the methylation and maintenance cost of the *E. coli* chemotaxis system: one full cycle of (de-)methylation requires (at least) 1 SAM, while the synthesis of a readout (CheY) or receptor (Tar/Tsr/Trg) protein requires about 10^4 ATPs. The ratio of the free energy drops associated with ATP and SAM hydrolysis is given by $\Delta\mu_{\text{SAM}}/\Delta\mu_{\text{ATP}} \approx 2$ [55]. The units of C are chosen such that they correspond to the number of proteins per hour. (B) Including the phosphorylation cost has a larger effect than including the methylation cost ($c_1 > 0$, $c_2 = 0$). Gray lines correspond to the scenario without phosphorylation cost, as in Fig. 4.3A. Inset: (C) Including both methylation and phosphorylation cost ($c_1, c_2 > 0$) combines the effects of panel A and B, and reduces the accessible region in the $(I_{\text{past}}, I_{\text{pred}})$ plane further. (D) The adaptation time over the signal correlation time, τ_{m}/τ_v as a function of the distance θ along the isocost lines in panel A (D-I), panel B (D-II), and panel C (D-III), gray lines correspond to the case without operating cost, Fig. 4.3B. For $\theta \to 0$, the methylation time τ_{m} goes to zero, reducing both I_{past} and I_{pred} to zero as the gain vanishes. Including only methylation cost marginally increases the optimal adaptation time under high resource availability (dashed line, D-I). Including only phosphorylation cost increases the optimal methylation time for any given C due to smaller resource availability for protein synthesis (D-II). Including both methylation and phosphorylation cost combines these effects (D-III). Other parameter values are (see Table 4.1): $g = 4\text{mm}^{-1}$, $\tau_{\text{r}} = 0.1\text{s}$, $\tau_v^{-1} = 0.9\text{s}^{-1}$, $K_{\text{D}}^{\text{I}} = 18\mu\text{M}$, $K_{\text{D}}^{\text{A}} = 2900\mu\text{M}$, $N = 12$, $\tilde{\alpha} = 2$, $p = 0.3$, $f = 0.5$, $\bar{\ell} = 100\mu\text{M}$, $\lambda^{-1} = 1\text{h}$, $c_1 = 10^{-4}/\Delta\mu_{\text{ATP}}$, $c_2 = 10^{-4}/\Delta\mu_{\text{SAM}}$. For the estimate of c_1, see Fig. 3.5.

elucidate the effects of the phosphorylation and methylation cost both separately and jointly, we consider three cases; i) c_1 is 0 and c_2 is positive (Fig. 4.4A and D-I), ii) c_1 is positive and c_2 is 0 (Fig. 4.4B and D-II), iii) both c_1 and c_2 are positive (Fig. 4.4C and D-III). These cases we compare against the result without operating cost (Fig. 4.3A and B), i.e. where both c_1 and c_2 are 0.

We find that the operating cost of the methylation cycle [55, 59] pushes the system further away from the information bound (Fig. 4.4A). Yet, for biologically realistic values of C, the effect is very small: while the free energy of SAM hydrolysis, driving the methylation cycle, is comparable to that of ATP hydrolysis, driving the phosphorylation cycle [70], the optimal methylation time $\tau_{\mathrm{m}}^{\mathrm{opt}}$ is about two orders of magnitude longer than the relaxation time τ_{r} of the phosphorylation cycle [13, 56]. Only for much higher values of C, when $\tau_{\mathrm{m}}^{\mathrm{opt}}$ becomes substantially shorter, does the methylation cost significantly move the system further away from the bound, reducing the accessible region of $(I_{\mathrm{pred}}, I_{\mathrm{past}})$ noticeably (Fig. 4.4A, D-I). Indeed, in the biologically relevant regime, the phosphorylation cost reduces this region more than the methylation cost (Fig. 4.4B).

4.5. COMPARISON WITH EXPERIMENT

To study how close the *E. coli* chemotaxis system comes to the information bound, we compute the predictive and past information directly from experimental data. Within the Gaussian framework, the predictive and past information can be obtained from the signal correlation function, noise correlation function, and network response kernel. Interestingly, these functions have recently been measured experimentally [20]. This indeed makes it possible to compute the predictive and past information from experimental data without the need to invoke a biochemical model. The only assumption required is that the system obeys Gaussian statistics, but recent work indicates that this is the case, to an excellent approximation [35].

Mattingly *et al.* measured the response kernel $K_b(t)$, the variance of the output noise σ_n^2, and the correlation function of the input $\langle s(t)s(t')\rangle$ to compute the information transmission rate between the encountered concentration derivative and the kinase activity [20]. It was found that the correlation function of the signal, i.e. the (relative) concentration derivative, is to a good approximation, given by

$$\langle s(t)s(t')\rangle \simeq g^2 a_v \exp(-\lambda|t-t'|), \tag{4.47}$$

where g is the gradient steepness, $a_v = 157.1 \pm 0.5\mu\mathrm{m}^2\mathrm{s}^{-2}$ is the variance in the positional derivative, and $\lambda = 0.862 \pm 0.005\mathrm{s}^{-1}$ is the correlation decay rate. The contribution of the noise in the output variance is $\sigma_n^2 = 0.092 \pm 0.002$. Finally, the response kernel was measured in experiments with 10μM step inputs in a 100μM background concentration to be

$$K_b(t) = G_b \exp(-t/\tau_2)(1-\exp(-t/\tau_1)). \tag{4.48}$$

The kernel gain was measured to be $G_b = 1.73 \pm 0.03$; the adaptation time $\tau_2 = 9.90 \pm 0.30$s, which agrees with earlier estimates (Table 4.2). The rise time was measured to be $\tau_1 = 0.22 \pm 0.01$s. However, the authors note that the rise time could be as short as 0.02s [20, 69, 71]. We use a rise time of $\tau_1 \approx 0.1$s, which lies well within formerly reported attractant and repellent response times [56], and corresponds to the measured time by Mattingly *et al.* minus the stimulus switching delay of ~ 0.1s induced by their microfluidic device.

Following Section 3.3, specifically using Eq. 3.6 and Eq. 3.17, we compute the past and predictive information between the current kinase activity and past or future signal from the measured response kernel $K_b(t)$, noise variance σ_n^2, and signal correlation function $\langle \delta s(0) \delta s(t) \rangle$, as follows:

$$I_{\rm past}^{\rm exp} = \frac{1}{2} \log\left(1 + \frac{\int_{-\infty}^{0} dt \int_{-\infty}^{0} dt' K_b(-t) K_b(-t') \left\langle s(t) s(t') \right\rangle}{\sigma_n^2}\right), \tag{4.49}$$

$$I_{\rm pred}^{\rm exp} = -\frac{1}{2} \log\left(1 - \frac{\left(\int_{-\infty}^{0} dt K_b(-t) \left\langle s(t) s(\tau) \right\rangle\right)^2}{g^2 a_v \left(\int_{-\infty}^{0} dt \int_{-\infty}^{0} dt' K_b(-t) K_b(-t') \left\langle s(t) s(t') \right\rangle + \sigma_n^2\right)}\right). \tag{4.50}$$

The resulting past and predictive information for cells swimming in an exponential concentration gradient are shown parametrically as a function of the gradient steepness g in Fig. 4.5 (blue line). It is seen that the *E. coli* chemotaxis system is remarkably far away from the information bound. This raises the question how close this system is to an optimal system that maximizes the predictive information under a resource constraint.

To compare the past and predictive information of *E. coli* to those of an optimized system, we fit the parameters of our model (Eqs. 4.17-4.19) to the experimental data of Mattingly *et al.* [20]. Before doing so, we need to consider that they use the kinase activity as output, which is determined from the FRET fluorescence of CheY_p bound to its phosphatase CheZ [20, 72]. This means that the past and predictive information computed from their experimental data using Eqs. 4.49 and 4.50 corresponds to the mutual information between the input and $\text{CheY}_\text{p} - \text{CheZ}$. However, in their experiments both CheY and CheZ are overexpressed, such that the number of readout molecules is much larger than the number of receptor clusters. For this reason, the phosphorylation noise is negligible and the mapping from $\text{CheY}_\text{p} - \text{CheZ}$ to CheY_p is therefore to a good approximation deterministic. A deterministic mapping leaves the mutual information unchanged: the mutual information between input and $\text{CheY}_\text{p} - \text{CheZ}$ thus becomes identical to that between the input and CheY_p, as computed in our model. Importantly, the response kernel of our model (Eq. 4.31) has the same mathematical form as that measured experimentally (Eq. 4.48); as a result, when the parameters of our model, including the methylation time, agree with those measured experimentally, the past and predictive information of our model (using Eqs. 4.39 and 3.6 for the past and Eqs. 4.44 and 3.17 for the predictive information) agree with those directly based on the experimental data, computed using Eqs. 4.49 and 4.50. Fixing the parameters as described in Appendix 4.C and optimizing over the methylation time τ_m, we can plot the past and predictive information of the optimal chemotaxis network parametrically as a function of the gradient steepness g (Fig. 4.5, red line).

Figure 4.5 shows that both in the optimal system (red line) and the chemotaxis system (blue line) I_pred and I_past increase as the gradient steepness g increases. A steeper gradient increases the signal strength, $\sigma_v^2 \sim g^2$ (Eq. 4.7), which increases the predictive and past information. Yet, the figure also shows that in the optimal system I_pred rises

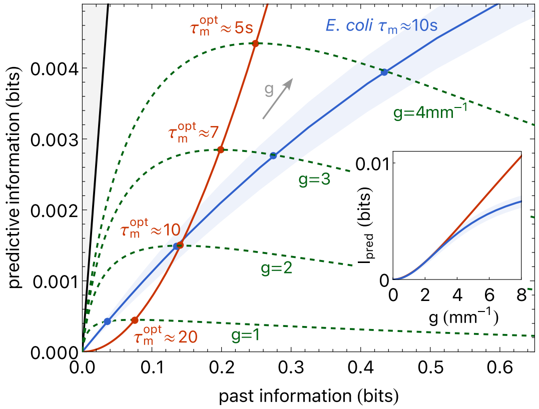


Figure 4.5: The *E. coli* chemotaxis network is far from the information bound, yet close to the optimal resource limited system in shallow gradients. Comparison of the *E. coli* chemotaxis system (blue line) to the optimal system (red line) as a function of the steepness g of an exponential ligand concentration gradient $\ell(x) = \bar{\ell}e^{gx}$. For the *E. coli* system, I_{pred} and I_{past} have been computed directly from the measured response kernel and signal and noise correlation functions [20] in a model-free manner. The optimal system is based on the model of Eqs. 4.17-4.19, in which the resource cost $C = NR_{\text{T}} + X_{\text{T}}$ is fixed, with $X_{\text{T}} = 10^4$, $N = 12$, and $R_{\text{T}} = 8$ as inferred from experimental data [20, 49], yet τ_{m} has been optimized; the dashed lines mark the boundaries of the accessible region of $(I_{\text{pred}}, I_{\text{past}})$ of this model upon varying τ_{m}, with the red dot corresponding to the system with the optimal τ_{m}. The predictive information increases with gradient steepness g, yet the optimal methylation time decreases with g. The inset shows that I_{pred} of the *E. coli* system is very close to that of the optimal system for shallow gradients $g \lesssim 4\text{mm}^{-1}$, but deviates from it at steeper gradients, suggesting the system has been optimized for sensing shallow gradients.

much faster with I_{past} than in the *E. coli* system. The stronger signal received in steeper concentration gradients raises the signal above the receptor sampling noise more, permitting a smaller gain. This allows the optimal system to take a more recent derivative, with a smaller τ_{m}, which is more informative about the future: the optimal methylation time scales as $\tau_{\text{m}}^{\text{opt}} \sim 1/g$ (Appendix 4.A, Fig. 4.6). In contrast, the methylation time τ_{m} of the *E. coli* chemotaxis system is fixed—the system cannot tune τ_{m} to g. As the inset in Fig. 4.5 shows, this methylation time is close to optimal for detecting shallow gradients, $g \lesssim 4\text{mm}^{-1}$: I_{pred} of the *E. coli* system is nearly identical to that of the optimal system. Moreover, in this regime, not only I_{pred} but also I_{past} is fairly similar for the two systems. For steeper gradients I_{past} becomes much higher in the *E. coli* system than in the optimal one, even though I_{pred} remains lower. The bacterium increasingly collects information that is less informative about the future. Taken together, these results strongly suggest that the system has been optimized to predict future concentration changes in shallow gradients, which necessitate a relatively long methylation time.

4.6. Discussion

In this chapter we have investigated whether biochemical networks exist which can reach the information bound for a generic class of non-Markovian input signals, corresponding to an underdamped particle in a harmonic well (Eqs. 4.1 and 4.2). To reach the bound for this class of input signals, the network should in general base its output on both the current signal value and its derivative. However, in the regime where the signal change over a run is much smaller than the range of background concentrations, as we argue is the regime in which *E. coli* operates, the current value is no longer correlated with the future derivative; only the current derivative is still correlated with the future derivative. In this regime, the optimal kernel is thus a perfectly adaptive kernel, in which the response only depends on the current derivative, but no longer on the current value of the concentration. Interestingly, this is the key characteristic of the kernel of *E. coli.* This result also shows that the signaling network of *E. coli* has the topology that makes it possible to reach the information bound.

However, as we also found for the push-pull network in the previous chapter, reaching the information bound is exceedingly costly. For the chemotaxis network, the reason is that taking an instantaneous derivative reduces the gain to zero, such that the signal is no longer lifted above the inevitable intrinsic biochemical noise of the signaling system. Moreover, the chemical power cost to drive the adaptation cycle [55, 59] pushes the system away from the information bound, although this effect is relatively small.

While the predictive information of the *E. coli* chemotaxis system in shallow gradients is very close to that of the optimal chemotaxis system with the same topology and resource constraint, it is much further away from the information bound than other information processing systems, in neuronal signaling [11] and embryonic development [73]. We believe this is because the chemotaxis systems take a temporal derivative only. This computation is inherently costly because it is based on signal subtraction, which dramatically reduces the gain. A small gain is particularly detrimental in the noise-dominated regime of low past and predictive information, where the signal strength σ_v^2 and the resource availability C are low. To maximize the predictive information in this regime, it becomes paramount to raise the gain by increasing the time interval over which the derivative is taken, as set by τ_{m}. However, this moves the system inevitably away from the bound.

Recent work has demonstrated that the chemotaxis receptors can exhibit strong synchronized switching [74], and evidence is mounting that receptor activity switching is a non-equilibrium process [75–77]. Moreover, even within an equilibrium MWC description, the receptor activity can become bistable, which could lead to oscillations on the timescale of methylation [78]. The model we employ here captures none of these features, raising the question how accurately it can quantify the information transmission in the *E. coli* chemotaxis network. While the description we use is clearly a simplified view of a much more complex system, it has been used before to successfully replicate a large part of the available experimental data [50, 52, 57, 58, 62, 66, 67, 79]. We can further obtain a model with an oscillatory regime by explicitly including the ligand binding dynamics in our model (see Appendix 4.B). However, with a background concentration of 100μM, as studied by Mattingly *et al.* [20] and used here throughout, the system is in a non-oscillatory regime that is well described by a linear model. Indeed, the integration kernel and correlation functions of our model closely match those that have been measured in recent experiments with the same background concentration [20]. Within a Gaussian description, the correlation functions fully determine the information transmission. How bistability in the receptor activity, oscillations on the methylation timescale, and possible non-equilibrium processes affect information transmission remain interesting questions which we leave for future work.

Any system that needs to predict the future environment, be it a living organism or a man-made device, must base its prediction on information it has collected from the past. In many cases, information from the recent past is likely to be more predictive than that from the distant past, as for the signals studied in this dissertation. It means that the system must continually and rapidly update its information on the environment. Yet, the laws of thermodynamics imply that there is a fundamental trade-off between precision, power, and speed [80]: obtaining information rapidly is inherently costly. We thus expect that our principal result, namely that there is a trade-off between obtaining information that is predictive versus that which is cheap, is generic, applying to a broad class of systems. For these systems, the optimal design that maximizes the predictive power under a resource constraint will differ from the design that maximizes predictive power under an information compression constraint.

Throughout this dissertation we define the past and predictive information as the mutual information between the past signal trajectory or a future signal property of interest and the current output value. However, information about the future state of the signal may be encoded in the full past output trajectory, rather than its instantaneous value only [46]. Yet, it is often not clear whether, nor how the cell decodes information stored in the past output trajectory. Indeed, the most natural information measure depends on the way in which the information encoded in the output is decoded, i.e. how the output determines the behavioral response of the cell. For example, the output of the *E. coli* chemotaxis network, CheY_p, is integrated over the intrinsic switching timescale of the flagellar motor before the motor switches direction. Therefore, the most natural information measure may in fact be the information stored in the CheY_p trajectory over the past motor integration time, rather than its instantaneous value or its entire past trajectory. However, because the motor simply averages CheY_p over a short time in the past—the motor integration time ($\sim 0.1-1$s [56, 81, 82]) is much shorter than the network's adaptation time (~ 10s [19, 20, 62])—the cell does not seem to have access to significantly more, or qualitatively different information than is contained in the instantaneous CheY_p concentration.

Information theory shows that the amount of transmitted information depends not only on the characteristics of the information processing system, but also on the statistics of the input signal. While much progress has been made in characterizing cellular signaling systems, the statistics of the input signal is typically not known, with a few notable exceptions [83]. Over the previous chapters, we have focused on two classes of input signals, but it seems likely that the signals encountered by natural systems are much more diverse. It will be interesting to extend our analysis to signals with a richer temporal structure [10], and see whether cellular systems exist that can optimally encode these signals for prediction.

Finally, while we have analyzed the design of cellular signaling networks to optimally predict future signals, we have not addressed the utility of information for function or behavior. It is clear that many functional or behavioral tasks, like chemotaxis [20], require information, but what the relevant bits of information are is poorly understood [7]. Moreover, cells ultimately employ their resources—protein copies, time, and energy—for function or behavior, not for processing information per se. Here, we have shown that maximizing predictive information under a resource constraint, $C \rightarrow I_{\text{past}} \rightarrow I_{\text{pred}}$, does not necessarily imply maximizing past information. This hints that optimizing a functional or behavioral task under a resource constraint, $C \rightarrow I_{\text{pred}} \rightarrow$ function, may not imply maximizing the predictive information necessary to carry out this task.

Appendices

4.A. Optimal design via signal to noise ratio

To maximize the predictive power of the chemotaxis network, its design, characterized by the ratio of readouts over receptor clusters $X_\mathrm{T}/R_\mathrm{T}$ and the adaptation time τ_m, must minimize the noise and maximize the part of the output which is informative on the future concentration change v_τ. The system must thus maximize the signal-to-noise ratio $\mathrm{SNR}_\mathrm{pred}$ (Eqs. 3.7-3.13), which is equivalent to minimizing Eq. 4.44, see Eq. 4.45.

Optimal design is independent of forecast interval. From Eqs. 4.43 and 4.44, in combination with Eq. 3.17, it is clear that the optimal design of the network that maximizes the predictive information does not depend on the forecast interval τ. This is because the dynamics of the signal, the derivative of the concentration, v, is Markovian. The forecast interval only affects the magnitude of the predictive information.

Noise. The noise $\mathrm{NOISE}_\mathrm{pred} = \sigma^2_{x|v_\tau}$ is given by $\sigma^2_{x|v_\tau} = \sigma^2_x - \tilde{g}^2\sigma^2_v$ (see Eqs. 3.13 and 3.9 with ℓ replaced by v). With the total variance σ^2_x given by Eqs. 4.36-4.38 and the dynamic gain $\tilde{g}$ given by Eq. 4.41, this yields:

$$\mathrm{NOISE}_\mathrm{pred} = \sigma^2_{x|L} + \sigma^2_{x|\eta} - \tilde{g}^2\sigma^2_v \tag{4.51}$$

$$\begin{aligned} =& X_\mathrm{T} f(1-f) + \bar{g}^2_{a\to x}\frac{1}{1+\tau_\mathrm{r}/\tau_\mathrm{m}}\alpha R_\mathrm{T} p(1-p) + \\ & \tilde{g}^2\sigma^2_v\left(e^{2\tau/\tau_v}\left(1+\frac{\tau_\mathrm{m}}{\tau_v}\right)\left(1+\frac{\tau_\mathrm{r}}{\tau_v}\right)\left(1+\frac{\tau_\mathrm{m}\tau_\mathrm{r}}{\tau_v(\tau_\mathrm{m}+\tau_\mathrm{r})}\right)-1\right). \end{aligned} \tag{4.52}$$

As for the push-pull network (Eq. 3.69), the total noise has three contributions. For the chemotaxis network they consist of: (1) readout (de)modification; (2) receptor methylation; (3) signal variations. The readout (de)modification noise is $X_\mathrm{T} f(1-f)$ and cannot be averaged out. The methylation noise at the level of the cluster activity is $\alpha R_\mathrm{T} p(1-p)$, but that propagated to the output x^* is the amount $\bar{g}^2_{a\to x}/(1+\tau_\mathrm{r}/\tau_\mathrm{m})R_\mathrm{T} p(1-p)$. This contribution can not be time averaged effectively, because for the chemotaxis network generally $\tau_\mathrm{m} > \tau_\mathrm{r}$. Since the static gain from the cluster activity to the readout $\bar{g}_{a\to x} = X_\mathrm{T} f(1-f)/(R_\mathrm{T} p)$ decreases with R_T, the contribution of the methylation noise to σ^2_x scales as $1/R_\mathrm{T}$. The third contribution, from the signal variations, increases with the adaptation time τ_m for two reasons. Firstly, a longer adaptation time τ_m increases the gain $\tilde{g}$ (the prefactor), which increases the contribution of both informative and uninformative signal fluctuations. Secondly, a longer adaptation time leads to an older derivative on average, which is less correlated with the future concentration derivative v_τ which the cell needs to predict.

Gain and Signal. The signal-to-noise ratio $\mathrm{SNR}_\mathrm{pred}$ is further set by the 'signal', which is the product of squared dynamic gain $\tilde{g}^2$ and the variance σ^2_v, see also Section 3.3. The dynamic gain $\tilde{g}$, given by Eq. 4.41, quantifies the degree to which variations in the output x^* are informative on the future concentration change v_τ. The dynamic gain depends on the static gain $\bar{g}_{v\to x} = \tau_\mathrm{m}\kappa N X_\mathrm{T} f(1-f)(1-p)$ (Eq. 4.33), such that the signal scales with X^2_T. Substituting the static gain $\bar{g}_{v\to x}$ in the dynamic gain $\tilde{g}$ (Eq. 4.41), shows that the gain increases with τ_m but saturates as $\tau_\mathrm{m} \gg \tau_v$ (see Eq. 4.42). The intuition is that input concentrations ℓ that are further apart in time are more dissimilar on average, such that subtracting them from one another increases the gain. The gain saturates when $\tau_{\mathrm{m}} \gg \tau_v$ (Eq. 4.42) because the signal that is subtracted from the recent signal becomes the average, baseline signal; in this regime, the system has essentially become a device that copies the most recent signal into the output, rather than one that takes a temporal derivative. Finally, $\tilde{g}$ decreases with the forecast interval τ

$\mathbf{SNR_{pred}}$. The relative error $\mathrm{SNR}^{-1}_{\mathrm{pred}}$ in predicting the future derivative is given by Eq. 4.45 with the instantaneous correlation coefficient given by Eq. 4.44. We discuss the relative error for prediction of the future derivative in more detail in Chapter 5. The first two terms on the right-hand side of Eq. 4.44 give the error arising from the readout modification noise and the receptor methylation noise. This error is analogous to the 'sampling error' of the push-pull network in [15], also see Appendix 3.D. Indeed, this sampling error also arises here because also in this chemotaxis system the push-pull network downstream of the receptor is a device that samples the receptor state in a stochastic and discrete fashion; the principal difference between the chemotaxis system and a bare push-pull network, as analyzed in Section 3.4, is that the chemotaxis system takes a temporal derivative at the receptor level. Importantly, the SNR depends on the noise and the gain squared, see Eq. 3.12. Hence, the sampling error decreases with the number of readout molecules as $1/X_{\mathrm{T}}$ because while the readout-modification noise increases with X_{T}, the gain-squared increases as X^2_{T}. While the receptor methylation noise increases with R_{T}, the gain from the input signal to the receptor goes as R^2_{T}, such that the signal-to-noise ratio scales as $1/R_{\mathrm{T}}$ (see for a more detailed explanation, Appendix 3.D). We thus see that increasing X_{T} and R_{T} raises the SNR by increasing the static gain, which helps to lift the signal above the readout-modification and receptor-methylation noise. The relative prediction error of the chemotaxis network further depends on its 'dynamical error' (Eq. 4.44), very similar to that in [15, 29]. This dynamical error is independent of X_{T} and R_{T} and only depends on ratios of timescales. In particular, this error increases with τ_{m}, because increasing τ_{m} means that the derivative is taken over a longer interval back into the past, which is less informative about the future derivative (which the cell aims to predict) than a more recent derivative taken with a shorter τ_{m}. This dynamical error is unique to the predictive information I_{pred}; for the past information I_{past}, the full variance $\sigma^2_{x|\eta}$ caused by fluctuations in the past ligand concentration is part of the 'signal' in the signal-to-noise ratio (Eqs. 3.6 and 4.39). The dynamical error only goes to zero when both the adaptation time τ_{m} and the integration time τ_{r} become much shorter than the derivative correlation time τ_v.

Optimal adaptation time. The optimal adaptation time $\tau^{\mathrm{opt}}_{\mathrm{m}}$ to predict the future concentration change is determined by a trade-off between mitigating the sampling error and the dynamical error, given by, respectively, the first two terms and the the final term in Eq. 4.44. The sampling error decreases monotonically with τ_{m} because a longer adaptation time increases the gain (Eq. 4.41), while the dynamical error increases monotonically with τ_{m} because a longer adaptation time leads to an output based on inputs further back into the past, which are less correlated with the future derivative. We discuss the optimal adaptation time in more detail in Chapter 5.

Scaling optimal adaptation time $\boldsymbol{\tau^{\mathrm{opt}}_{\mathrm{m}}}$ with resource cost $\boldsymbol{C}$ and gradient steepness $\boldsymbol{g}$. To obtain an insightful analytical result we consider that the phosphorylation time τ_{r} is much shorter than both the methylation time τ_{m} and the input timescale τ_v, $\tau_{\mathrm{r}} \ll$

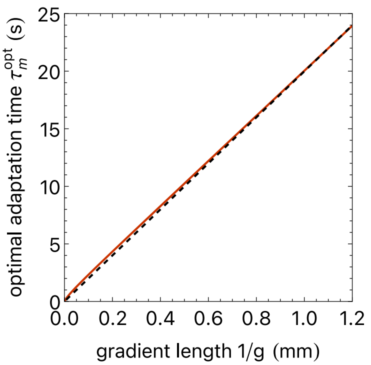


Figure 4.6: The optimal methylation time as a function of the gradient length. The optimal methylation time (red solid curve) has been determined numerically by minimizing the prediction error (Eq. 4.44). It scales to an excellent approximation linearly with the gradient length $1/g$, as illustrated by the black dashed line, which gives the analytical result of Eq. 4.54. Parameters: $R_{\rm T}=8$, $N=12$, $X_{\rm T}=10^4$, $\tau_{\rm r}=0.1$s, $\tau_v^{-1}=0.9{\rm s}^{-1}$, $K_{\rm D}^{\rm I}=18\mu$M, $K_{\rm D}^{\rm A}=2900\mu$M, $\tilde{\alpha}=2$, $p=0.3$, $f=0.5$, $\bar{\ell}=100\mu$M.

$\tau_{\rm m}, \tau_v$. In this regime, the derivative of the relative error (Eq. 4.44) with respect to the adaptation time becomes

$$\frac{\partial \rho_{vx}^{-2}}{\partial \tau_{\rm m}} \approx \frac{1}{\tau_v} - \frac{2(1+\tau_{\rm m}/\tau_v)}{\tau_{\rm m}^3 \sigma_v^2 \kappa^2 N^2 C}\left(\frac{1}{\phi_X f(1-f)(1-p)^2} + \frac{\tilde{\alpha} N}{\phi_R}\right) \qquad \text{for } \tau_{\rm r} \ll \tau_v, \tau_{\rm r} \ll \tau_{\rm m}, \quad (4.53)$$

where $\phi_X \equiv X_{\rm T}/C$ and $\phi_R \equiv R_{\rm T}/C$. A clear scaling of $\tau_{\rm m}^{\rm opt}$ with the resource availability C and signal variance σ_v^2 can be found when considering the biologically relevant regime of shallow gradients. In shallow gradients the signal variance is small because $\sigma_v^2 \propto g^2$ (see Eq. 4.7). Because the signal is weak (i.e. signal variance is small), the adaptation time must be relatively large in order to lift the signal above the noise, i.e. reduce the sampling error (Eq. 4.44): this means $\tau_{\rm m} > \tau_v$ (see also Fig. 4.5). In this regime the optimal methylation time (obtained by setting the derivative of Eq. 4.53 to zero) becomes

$$\tau_{\rm m}^{\rm opt} \approx \sqrt{\frac{2}{\sigma_v^2 \kappa^2 N^2 C}\left(\frac{1}{\phi_X f(1-f)(1-p)^2} + \frac{\tilde{\alpha} N}{\phi_R}\right)} \qquad \text{for } \tau_{\rm m} \gg \tau_v \gg \tau_{\rm r}. \qquad (4.54)$$

Since ϕ_X and ϕ_R vary only weakly with C (see Fig. 4.7), we see that the optimal methylation time $\tau_{\rm m}^{\rm opt}$ scales as $1/\sqrt{C}$. Moreover, since $\sigma_v^2 \propto g^2$ (Eq. 4.7), in the shallow gradient regime the optimal adaptation time is proportional to $1/g$, i.e. the gradient length. We test the result of Eq. 4.54 by numerically finding the optimal $\tau_{\rm m}$ that minimizes the prediction error (Eq. 4.44) for a range of gradient lengths $1/g$, with all other model parameters as inferred from experimental data (Fig. 4.6). The result shows that Eq. 4.54 is highly accurate, even as the gradient length $1/g$ decreases, i.e. the gradient steepness g increases.

Optimal ratio $X_{\rm T}/R_{\rm T}$. We can determine the optimal ratio $(X_{\rm T}/R_{\rm T})^{\rm opt}$ that maximizes both the past information and the predictive information, given all the other network

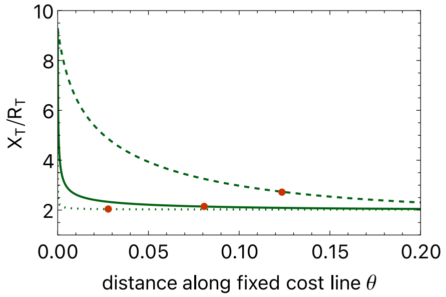


Figure 4.7: The optimal allocation ratio X_T/R_T varies only weakly along the isocost lines of Fig. 4.3 in the main text. The optimal ratio X_T/R_T as a function of the distance θ along the isocost lines for $C = NR_T + X_T$ of Fig. 4.3A of the main text; dotted line $C = 10^2$, solid line $C = 10^4$, dashed line $C = 10^6$. The red dots mark the points where the predictive information is maximal. Along the isocost lines X_T/R_T varies much more weakly than for the push-pull network (Fig. 3.3B); for resource availability $C \leq 10^4$ the ratio is almost constant. Parameters used $N = 12$, $g = 4\text{mm}^{-1}$, $\tau_r = 0.1\text{s}$, $\tau_v^{-1} = 0.9\text{s}^{-1}$, $K_D^I = 18\mu\text{M}$, $K_D^A = 2900\mu\text{M}$, $\tilde{\alpha} = 2$, $p = 0.3$, $f = 0.5$, $\bar{\ell} = 100\mu\text{M}$.

parameters, most notably τ_m. Just as for the push-pull network, the number of readouts and receptors and their ratio only affects the sampling error Eq. 4.44. Therefore, the optimal ratio $(X_T/R_T)^{\text{opt}}$ is the same regardless of whether the past or the predictive information is maximized (see Eqs. 4.39 and 4.44). This is because the information on the future signal (be it the value or the derivative) is encoded in the receptor occupancy, and the ratio X_T/R_T, together with τ_r, controls the interval by which the downstream readout samples the receptor state to estimate its occupancy [15]. Nonetheless, the optimal methylation timescale τ_m^{opt} that maximizes either the past or the predictive information is different—maximizing predictive information requires a more recent derivative and hence a shorter τ_m than obtaining past information.

Given τ_m and all other parameters, the optimal ratio of the number of readout molecules over receptor clusters can be found by expressing X_T and R_T in terms of X_T/R_T using $C = AR_T + BX_T$ (also see Eqs. 3.70 and 3.71), taking the derivative of Eq. 4.44 with respect to X_T/R_T and equating it to zero. This yields

$$\left(\frac{X_T}{R_T}\right)^{\text{opt}} = \sqrt{\frac{A}{\alpha B}\frac{1}{f(1-f)}\frac{p}{1-p}}\sqrt{1+\frac{\tau_r}{\tau_m}}, \tag{4.55}$$

where $\alpha = \tilde{\alpha}Np(1-p)$ (see Eq. 4.17), and the optimal allocation ratio is thus independent of the cluster size for $A = N$. Because for the chemotaxis network $\tau_r < \tau_m$ the ratio τ_r/τ_m only varies between 0 and 1. For this reason, the optimal ratio $(X_T/R_T)^{\text{opt}}$ depends only weakly on τ_m, and does not vary strongly along the isocost lines of Fig. 4.3A in the main text, see Fig. 4.7.

4.B. CHEMOTAXIS MODEL WITH LIGAND BINDING

In the main text, we integrated out the ligand binding dynamics in our chemotaxis model based on the assumption that ligand binding is fast compared to all other timescales of the network. Here, we will investigate whether this assumption has a significant effect on the behavior of the network by studying a chemotaxis model which explicitly includes the ligand binding dynamics. We will show that the integration kernel and the power spectra of the network do not significantly differ from those in the main text. However, we find that a positive feedback arises between cluster activity and ligand binding, which tends to enhance the gain; to reach the experimentally observed gain, a smaller cluster size would be needed.

We start by again defining our chemotaxis model and perform a linear approximation of its dynamics, this time explicitly including ligand binding dynamics. We then study the stability of the linear model, which depends on the interplay between ligand binding and receptor cluster activity. Finally, we derive the integration kernel and power spectra of the network.

As in the main text, we approximate the probability of a receptor cluster to be active by its equilibrium probability, given by

$$a(n,m) = \frac{1}{1+\exp[\beta\Delta F(n,m,t)]} \tag{4.56}$$

with the difference in free energy of the active and inactive cluster state $\Delta F = F_A - F_I$. The free energy of the cluster state depends on the number of bound ligands n and methylated sites m. We have

$$\Delta F(n,m,t) = \Delta F_0 + (\Delta G_A - \Delta G_I)\, n(t) - (\mu_B + \mu_R)\, m(t), \tag{4.57}$$

The free energy difference consists of the free energy difference of the basal state $\Delta F_0 = F_0^A - F_0^I$, the difference in the free energy changes associated with ligand binding ΔG_A and ΔG_I, obeying $0 > \Delta G_A > \Delta G_I$, and the driving costs of methylation μ_R and of demethylation μ_B (both larger than 0).

The free energy change due to ligand binding is related to the dissociation constant of the binding reaction. The thermodynamic equilibrium constant of a reaction involving r species is defined as follows

$$K \equiv \prod_i^r \left(\frac{c_i}{c_0}\right)^{s_i} = e^{-\beta\Delta G}. \tag{4.58}$$

In this definition c_0 is a reference concentration and c_i is the concentration of species i, the stoichiometric coefficient of this species is s_i, and ΔG is the change in Gibbs free energy associated with the reaction. For the reaction $R + L \rightleftharpoons RL$ the dissociation constant of ligand binding is defined as $K_D \equiv c_R c_L / c_{RL}$. Using Eq. 4.58 we obtain $K = (c_0 c_{RL})/(c_R c_L)$ thus giving

$$K_D = c_0/K = c_0 e^{\beta\Delta G}, \tag{4.59}$$

$$\Delta G = \beta^{-1}\ln(K_D/c_0). \tag{4.60}$$

Using this relation and Eq. 4.57 we express Eq. 4.56 in terms of n and m:

$$a(t) = \frac{1}{1+\exp[\beta\Delta F_0 + \ln(K_D^A/K_D^I)n(t) - \tilde{\alpha} m(t)]}, \tag{4.61}$$

with $\tilde{\alpha} = \beta(\mu_R + \mu_B)$, and the dissociation constants of the active and inactive receptor states, K_D^A and K_D^I respectively.

Dynamics

The number of methylated sites on a receptor cluster depends on its activity state. We assume that methylation operates in a linear regime and that the concentrations of the (de-)methylation enzymes CheR and CheB are saturating, such that we can absorb them in the reaction rates. We effectively study the system

$$I_m + Q^* \underset{k_R^-}{\overset{k_R^+}{\rightleftharpoons}} I_{m+1} + Q, \tag{4.62}$$

$$A_m \underset{k_B^-}{\overset{k_B^+}{\rightleftharpoons}} A_{m-1} + q, \tag{4.63}$$

where I_m and A_m respectively are inactive and active receptor clusters with m methylated sites, Q^* methylates the inactive receptor, and q is the released methyl group after demethylation of the active receptor. As mentioned above, both the methylation and demethylation reaction are driven out of equilibrium. We consider the concentration of methylation species Q^*, Q and q to be fixed, and obtain in the irreversible limit

$$\dot{m}_i = (1 - a_i(n_i, m_i))k_R - a_i(n_i, m_i)k_B + B_{m_i}(a_i)\xi(t), \tag{4.64}$$

These dynamics indeed give rise to perfect adaptation, as setting the deterministic part to zero we find for the steady state cluster activity

$$p \equiv a_i(\bar{n}, \bar{m}) = \frac{1}{1 + k_B/k_R}, \tag{4.65}$$

which is independent of the external ligand concentration. When the dynamics would be reversible the system is still adaptive, but with an error [59].

The dynamics of the number of bound ligands on a cluster are

$$\dot{n}_i = k_+(N - n_i(t))\ell(t) - n_i(t)k_+\left(a_i(t)K_D^A + (1 - a_i(t))K_D^I\right) + B_{c_i}(a_i, n_i)\xi_c(t), \tag{4.66}$$

with the number of receptors per cluster N, and assuming that the binding rate is the same for both the active and inactive state of the receptor. We find for the steady state fraction of bound ligands to the cluster

$$\phi \equiv \bar{n}/N = \frac{\bar{\ell}}{\bar{\ell} + pK_D^A + (1-p)K_D^I}. \tag{4.67}$$

Active receptor clusters catalyze the phosphorylation of read-out molecules and phosphorylated read-out molecules decay at a constant rate. We have

$$\dot{x}^* = \sum_{i=1}^{R_T} a_i(t)(X_T - x^*(t))k_f - x^*(t)k_r + B_x(a_i, x)\xi_x(t), \tag{4.68}$$

where R_T is the total number of receptor *clusters*, rather than the total number of receptors as in Chapter 3. The steady state fraction of phosphorylated read-outs is now given by

$$f \equiv \bar{x}^*/X_T = \frac{R_T p}{R_T p + k_r/k_f}. \tag{4.69}$$

In the linear approximation we obtain for the deviation of a cluster's activity from its steady state value $\delta a_i(t) = a_i(t) - p$,

$$\delta a_i(t) = \alpha\delta m_i(t) - \kappa\delta n_i(t), \tag{4.70}$$

With $\alpha = p(1-p)\tilde{\alpha}$ and $\kappa = p(1-p)\ln(K_D^A/K_D^I)$. The factors $p(1-p)$ arise from the partial derivatives of $a_i(n_i, m_i)$. Directly implementing Eq. 4.70 in the linearization of the binding, methylation, and phosphorylation dynamics we obtain the complete dynamics of the system:

$$\frac{d\sum_i^{R_T}\delta n_i}{dt} = b\delta\ell(t) - \sum_i^{R_T}\Big((1-\theta)\delta n_i(t) + \frac{\alpha}{\kappa}\theta\delta m_i(t)\Big)/\tau_c + B_c\xi_c(t), \tag{4.71}$$

$$\frac{d\sum_i^{R_T}\delta m_i}{dt} = \sum_i^{R_T}\Big(\frac{\kappa}{\alpha}\delta n_i(t) - \delta m_i(t)\Big)/\tau_m + B_m\xi_m(t), \tag{4.72}$$

$$\frac{d\delta x^*}{dt} = -\delta x^*(t)/\tau_r + \gamma\sum_{i=1}^{R_T}(\alpha\delta m_i(t) - \kappa\delta n_i(t)) + B_x\xi_x(t). \tag{4.73}$$

We have written the dynamics in terms of the sum of bound ligands and methylated sites over all clusters, as this is what eventually determines the number of phosphorylated read-outs. We have the 'naive' (also see the following section) decorrelation rates

$$\tau_c^{-1} = k_+(\bar{\ell} + pK_D^A + (1-p)K_D^I), \tag{4.74}$$

$$\tau_m^{-1} = \alpha(k_R + k_B), \tag{4.75}$$

$$\tau_r^{-1} = R_T p k_f + k_r. \tag{4.76}$$

Further parameters are the ligand binding rate b upon a deviation $\delta\ell(t)$, the read-out phosphorylation rate γ upon a deviation $\delta a(t)$, and the relative positive feedback strength θ between ligand binding and receptor activity:

$$b = (1-\phi)R_T N k_+ = \phi(1-\phi)R_T N/(\bar{\ell}\tau_c), \tag{4.77}$$

$$\gamma = f(1-f)X_T/(pR_T\tau_r), \tag{4.78}$$

$$\theta = \kappa\phi N(K_D^A - K_D^I)k_+\tau_c. \tag{4.79}$$

The positive feedback between ligand binding and receptor activity arises because ligand binding reduces receptor activity, which, via the necessary change in the dissociation constant due to detailed balance, in turn enhances ligand binding. The strength of the positive feedback depends on the change in activity upon ligand binding κ (defined below Eq. 4.70) and the number of bound ligands ϕN, but most notably on the difference between the dissociation constants of the active and inactive cluster state $K_D^A - K_D^I$. The feedback strength is measured relative to τ_c^{-1}, note that $k_+\tau_c = (\bar{\ell} + pK_D^A + (1-p)K_D^I)^{-1}$, and the positive feedback is thus independent of the on rate k_+. For the linear system to be stable we require $\theta < 1$ (also see the following section). Using these parameter definitions we obtain for the noise strengths in the sums over all clusters

$$B_c = \sqrt{2R_T N\phi(1-\phi)/\tau_c}, \tag{4.80}$$

$$B_m = \sqrt{2R_T p(1-p)/(\alpha\tau_m)}, \tag{4.81}$$

$$B_x = \sqrt{2X_T f(1-f)/\tau_r}. \tag{4.82}$$

STABILITY OF CLUSTER ACTIVITY

Above we have performed a linear approximation of the full dynamics of the chemotaxis network including ligand binding [30]. This procedure often allows for a good approximation of dynamical systems that fluctuate around a single steady state. However, both theoretical and experimental work indicate that the receptor cluster can become bistable [74, 76, 78]. In the equilibrium description that we use here, this bistability can arise too, from the positive feedback between ligand binding and receptor activity: as more ligand molecules bind a receptor cluster it becomes less active, leading to increased ligand binding of the other receptors in the cluster (and vice versa). In the bistable regime, the positive feedback leads to receptor clusters being preferably completely bound or completely unbound. At the same time, the active (de)methylation of the receptors pushes the cluster activity back towards an intermediate fixed point. Together, these two effects—bistability of the cluster activity and adaptation towards a fixed activity level—can lead to oscillations in the cluster activity [78]. This oscillatory regime of the full nonlinear model corresponds to an unstable regime of the linear model. Therefore, to verify that we expand around a single, stable steady state, we investigate the stability of the linear model.

The stability of the linear model (Eqs. 4.70-4.73) depends on the positive feedback parameter θ (Eq. 4.79), which is primarily determined by the number of receptors per cluster N, the difference between the dissociation constants, and the steady state fraction of ligand bound receptors per cluster ϕ. We consider the first two of these to be fixed, but ϕ is determined by the background concentration, which depends on the environment. We can obtain a specific condition on θ by inspecting the stability of the cluster activity in the linear approximation. The Jacobian of the activity is given by,

$$\boldsymbol{J}_a = \begin{pmatrix} -(1-\theta)/\tau_c & -\alpha\theta/(\kappa\tau_c) \\ \kappa/(\alpha\tau_m) & -1/\tau_m \end{pmatrix}, \tag{4.83}$$

which has eigenvalues

$$\lambda_1 = -\lambda_c = -(\mu_a + \nu_a), \tag{4.84}$$

$$\lambda_2 = -\lambda_m = -(\mu_a - \nu_a), \tag{4.85}$$

with the mean of the receptor correlation and methylation timescales $\mu_a = ((1-\theta)\tau_c^{-1} + \tau_m^{-1})/2$ and $\nu_a = \sqrt{\mu_a^2 - (\tau_m\tau_c)^{-1}}$. Note that these eigenvalues give the effective relaxation timescales for receptor binding and adaptation. For $\theta \to 0$ we indeed find $\lambda_c, \lambda_m \to \tau_c^{-1}, \tau_m^{-1}$. For imaginary ν_a the linear system will oscillate, when μ_a becomes negative the linear system is unstable. We thus require for the system to be stable

$$\begin{aligned} &\mu_a > 0, \\ &\theta < 1 + \tau_c/\tau_m \approx 1, \end{aligned} \tag{4.86}$$

which holds for both small and large background concentrations $\bar{\ell}$, as shown in Fig. 4.8. For a background concentration of 100μM, as used in the experiments of Mattingly *et al.* [20] and throughout this work, the linearized system is stable (Fig. 4.8). This also suggests that for this background concentration, the full nonlinear system does not oscillate. Nonetheless, our analysis reveals that for higher concentrations, the fixed point becomes linearly unstable (Fig. 4.8), corresponding to the oscillatory regime of the nonlinear receptor dynamics. We emphasize however that while according to our analysis the individual receptor clusters will oscillate in this regime, the total activity of the receptor array, given by the sum of the activities of all receptor clusters, may not oscillate. The reason is that global oscillations only arise when the cycles of the individual clusters are synchronized. While limiting amounts of CheR and CheB may provide such a synchronization mechanism [84, 85], future work is needed to investigate this regime.

We continue by using the same background concentration as in the main text $\bar{\ell}$ = 100μM, for which the system is stable. In the section that follows, we will show how the model with explicit ligand binding can be mapped onto the model of the main text, where ligand binding is integrated out.

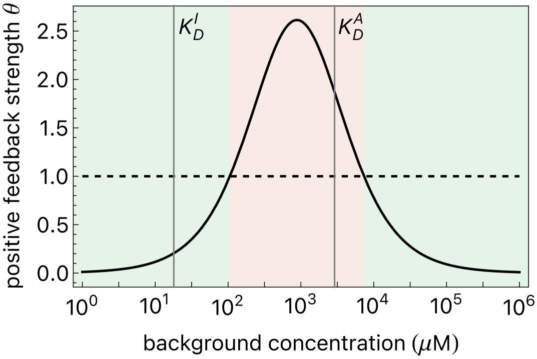


Figure 4.8: The stability of the linear chemotaxis model depends on the background concentration. Positive feedback strength θ as a function of the background concentration $\bar{\ell}$. When the positive feedback strength $\theta > 1$ the linear model is unstable, indicating that the full nonlinear model is in the oscillatory regime. A background concentration of 100μM, which we use throughout this work, lies (just) within the stable regime. Parameters: $p = 0.3$, $K_D^I = 18\mu$M, $K_D^A = 2900\mu$M, $N = 3$, see also the following section.

INTEGRATION KERNEL AND SPECTRAL PROPERTIES

To understand how the model discussed in the previous sections compares to that of the main text, we will here derive the integration kernel and the power spectra of the chemotaxis network including ligand binding dynamics. We follow the same procedure to derive the integration kernel and power spectrum as we did in the main text Section 4.2 (also see Appendix 3.C). The input signal (Eqs. 4.1 and 4.2) is characterized by both its concentration and derivative, its corresponding power spectra are

$$\mathbb{S}_s(\omega) = \begin{pmatrix} S_\ell(\omega) & i\omega S_\ell(\omega) \\ -i\omega S_\ell(\omega) & \omega^2 S_\ell(\omega) \end{pmatrix}, \tag{4.87}$$

$$S_\ell(\omega) = \frac{2\mu\sigma_v^2}{(\omega^2+(\mu/2+\rho)^2)(\omega^2+(\mu/2-\rho)^2)}, \tag{4.88}$$

with $\rho = \sqrt{\mu^2/4-\omega_0^2}$. The network is fully determined by the following matrices (Eq. 3.62)

$$\boldsymbol{G} = \begin{pmatrix} b & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \tag{4.89}$$

$$\boldsymbol{J} = \begin{pmatrix} -(1-\theta)/\tau_c & -\alpha\theta/(\kappa\tau_c) & 0 \\ \kappa/(\alpha\tau_m) & -1/\tau_m & 0 \\ -\gamma\kappa & \gamma\alpha & -1/\tau_r \end{pmatrix}, \tag{4.90}$$

$$\boldsymbol{B} = \begin{pmatrix} \sqrt{2R_T N\phi(1-\phi)/\tau_c} & 0 & 0 \\ 0 & \sqrt{2R_T p(1-p)/(\alpha\tau_m)} & 0 \\ 0 & 0 & \sqrt{2X_T f(1-f)/\tau_r} \end{pmatrix}. \tag{4.91}$$

the matrix $\boldsymbol{G}$ reflects that the network only has access to the ligand concentration. The Fourier transform of the matrix exponential of the Jacobian is

$$\begin{aligned} \mathcal{F}\{e^{\boldsymbol{J}t}\} &= (i\omega\mathbb{I}_3 - \boldsymbol{J})^{-1}, \\ &= \begin{pmatrix} \frac{1/\tau_m+i\omega}{(\lambda_c+i\omega)(\lambda_m+i\omega)} & \frac{-\alpha\theta/(\kappa\tau_c)}{(\lambda_c+i\omega)(\lambda_m+i\omega)} & 0 \\ \frac{\kappa/(\alpha\tau_m)}{(\lambda_c+i\omega)(\lambda_m+i\omega)} & \frac{(1-\theta)/\tau_c+i\omega}{(\lambda_c+i\omega)(\lambda_m+i\omega)} & 0 \\ \frac{-i\gamma\kappa\omega}{(\lambda_c+i\omega)(\lambda_m+i\omega)(1/\tau_r+i\omega)} & \frac{\alpha\gamma(1/\tau_c+i\omega)}{(\lambda_c+i\omega)(\lambda_m+i\omega)(1/\tau_r+i\omega)} & \frac{1}{1/\tau_r+i\omega} \end{pmatrix}. \end{aligned} \tag{4.92}$$

Here we have used the effective relaxation rates λ_c and λ_m defined in Eq. 4.84 and Eq. 4.85. The frequency dependent gain matrix is now given by $\mathbb{K}(\omega) = \mathcal{F}\{e^{\boldsymbol{J}t}\}\boldsymbol{G}$, and the noise matrix by $\mathbb{N}(\omega) = \mathcal{F}\{e^{\boldsymbol{J}t}\}\boldsymbol{B}$, also see Eq. 3.65.

The integration kernel from ligand concentration to output is given by the inverse Fourier transform of entry $(1,3)$ of the gain matrix $\mathbb{K}(\omega)$, which is

$$k(t) = \mathcal{F}^{-1}\left\{\frac{-ib\gamma\kappa\omega}{(\lambda_c+i\omega)(\lambda_m+i\omega)(1/\tau_r+i\omega)}\right\} \tag{4.93}$$

$$-k(t) = b\gamma\kappa\left(\frac{\frac{1}{\tau_r}e^{-t/\tau_r}}{(\lambda_c-1/\tau_r)(1/\tau_r-\lambda_m)} - \frac{\lambda_m e^{-\lambda_m t}}{(\lambda_c-\lambda_m)(1/\tau_r-\lambda_m)} - \frac{\lambda_c e^{-\lambda_c t}}{(\lambda_c-\lambda_m)(\lambda_c-1/\tau_r)}\right). \tag{4.94}$$

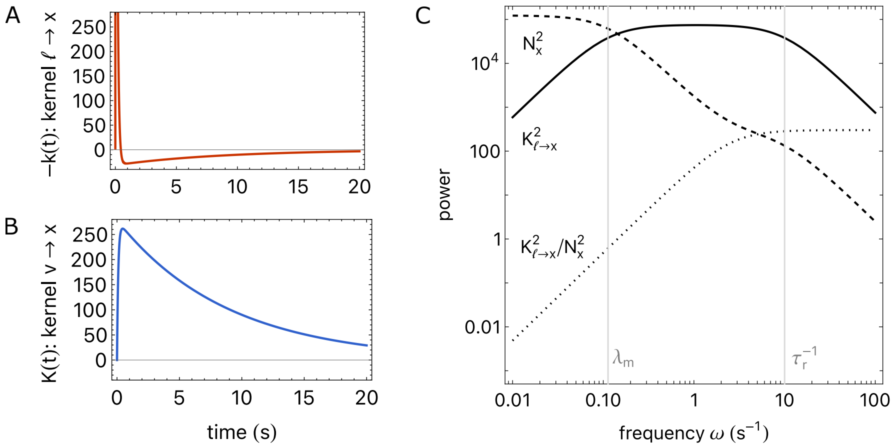


Figure 4.9: The integration kernel and power spectra of the chemotaxis model including ligand binding dynamics. These kernels and spectral properties are highly similar to those of the model in the main text where ligand binding is integrated out, see Fig. 4.2. Notably, the model considered here has a larger gain, $K^2_{\ell\to x}(\omega)$, and slightly larger noise, $N_x^2(\omega)$, and has a higher gain-to-noise ratio than the model in the main text for the same resource availability $R_T = X_T = 5000$. Timescales are $\tau_r = 0.1$s and $\tau_m = 200$s, where the latter is chosen such that the effective adaptation time is as measured experimentally $\lambda_m^{-1} \approx 10$s. Model parameters are $k_+ = 10^3\mu\text{M}^{-1}\text{s}^{-1}$, $\tilde{\alpha} = 2$, $N = 3$, $K_D^I = 18\mu$M, $K_D^A = 2900\mu$M, $\bar{\ell} = 100\mu$M, $p = 0.3$, $f = 0.5$.

This integration kernel is shown in Fig. 4.9A. As derived in the main text (Eqs. 4.30-4.32), the integration kernel from ligand concentration derivative to output is the primitive of the kernel $k(t)$ from ligand concentration to output; it is shown in Fig. 4.9B.

Before we discuss and compare the integration kernels to those of the model without ligand binding in the main text, we first determine the power spectrum of the readout x^* (via Eq. 3.66 with Eqs. 4.89-4.92),

$$S_x(\omega) = |K_{\ell\to x}(\omega)|^2 S_\ell(\omega) + |N_x(\omega)|^2,$$

$$= \frac{b^2\gamma^2\kappa^2\omega^2}{(\lambda_c^2+\omega^2)(\lambda_m^2+\omega^2)(1/\tau_r^2+\omega^2)} S_\ell(\omega) + \gamma^2 \frac{\kappa^2\omega^2 B_c^2 + \alpha^2(1/\tau_c^2+\omega^2)B_m^2}{(\lambda_c^2+\omega^2)(\lambda_m^2+\omega^2)(1/\tau_r^2+\omega^2)} + \frac{B_x^2}{1/\tau_r^2+\omega^2}. \tag{4.95}$$

The frequency dependent gain, noise, and their ratio are shown in Fig. 4.9C.

To compare the chemotaxis model discussed here, where ligand binding dynamics are explicitly included, to that of the main text, where they are integrated out, we need to determine the correct mapping of parameters between the different models. Firstly, the experimentally measured adaptation time of *E. coli*, which is approximately 10s, should here correspond to the effective adaptation time λ_m^{-1} rather than the underlying methylation timescale τ_m (see also Eq. 4.85). We find that an effective adaptation time of $\lambda_m^{-1} \approx 10$s here requires a very large methylation timescale $\tau_m \approx 200$s, while in the main text the methylation and adaptation time are equivalent. Moreover, in the main text we obtained the number of receptors per cluster N by fitting the gain of the kernel $K(t)$ that was measured experimentally [20]. Following the same procedure as discussed in the main text Section 4.5 and Appendix 4.C, we here obtain $N \approx 3$ instead of $N \approx 12$. The reason is that the positive feedback that arises between ligand binding and cluster activity amplifies the gain of the network. We thus require a smaller cluster size N to obtain the same (normalized) gain. This is also reflected in Fig. 4.9, where in comparison to Fig. 4.2 in the main text we find that the (absolute) gain of the network as we model it here is larger for the same resource availability $R_{\mathrm{T}} = X_{\mathrm{T}} = 5000$.

Apart from the gain, comparing Figs. 4.9 and 4.2 reveals that neither the power spectra nor the integration kernels changed significantly by including the ligand binding dynamics explicitly. The reason is that the timescale of ligand binding is much faster than that of methylation and phosphorylation. In Fig. 4.9A we see that the integration kernel now starts at 0, but it rises very quickly to its maximum. Therefore, the overall shape of the kernel remains very similar. From Fig. 4.9C we see that including ligand binding also does not qualitatively change the spectral properties of the network, as it still acts as a bandpass filter between the adaptation time, now λ_m^{-1}, and the integration time τ_{r}. Ligand binding only has an effect at much higher frequencies, which are not important for signal transmission, because the signal has a characteristic timescale of $\tau_v \approx 1\mathrm{s}$.

4.C. FITTING MODEL PARAMETERS TO EXPERIMENTAL DATA

Here we explain how we set the parameters of our model in the main text (Eqs. 4.17-4.19), to correspond to the parameter values measured by Mattingly *et al.* [20].

We start by fitting the gain of our model to that measured experimentally, $G_b \approx 1.73$ (see Eq. 4.48). The gain of our model, i.e. the amplitude of the kernel, is $G_a = \kappa X_{\mathrm{T}} N f(1-f)(1-p)/(1-\tau_{\mathrm{r}}/\tau_{\mathrm{m}})$ (see Eq. 4.31). Importantly, however, the output in the experiments of [20] is normalized, which in our model can be captured by normalizing its output, the number of phosphorylated readout molecules x^*, by the total number of readout molecules X_{T}. In addition, in [20] the signal is defined as the change in the log of the concentration, but in the regime of shallow gradients this can be captured by rescaling the response kernel by $\bar{\ell}$. We therefore divide the gain of our model by X_{T} and multiply it by $\bar{\ell}$, such that $\kappa\bar{\ell}Nf(1-f)(1-p)/(1-\tau_{\mathrm{r}}/\tau_{\mathrm{m}}) \approx 1.73$. Most of the microscopic parameters in this expression have been measured separately in the same work; $\bar{\ell} = 100\mu\mathrm{M}$, $p = 0.3$, $\tau_{\mathrm{m}} = \tau_2 = 9.9\mathrm{s}$ and $\tau_{\mathrm{r}} = (1/\tau_1 + 1/\tau_2)^{-1} \approx \tau_1 = 0.1\mathrm{s}$ [20]. We further have $\kappa = (\bar{\ell} + K_{\mathrm{D}}^{\mathrm{I}})^{-1} - (\bar{\ell} + K_{\mathrm{D}}^{\mathrm{A}})^{-1} = 0.008\mu\mathrm{M}^{-1}$ based on previously measured dissociation constants (Table 4.1). We assume the fraction of phosphorylated readouts in steady state $f = 0.5$. Fitting the gain then gives the only free parameter that is left, $N \approx 12$. This estimate appears reasonable: Early estimates of the cluster size were based on bulk dose-response measurements with a relatively slow ligand exchange, yielding $N \approx 6$ [57, 62]. More recent dose-response measurements, at the single cell level and with faster ligand exchange, yield an average that is higher $\langle N \rangle \approx 8$, and with a broad distribution around it, arising from cell-to-cell variability [67]. Furthermore, our estimate is close to a fit of $N \approx 15$ made of the same data using a nonlinear microscopic model [35].

Exploiting that CheY is overexpressed ($X_{\mathrm{T}} > R_{\mathrm{T}}$) and that the phosphorylation noise is therefore small compared to the methylation noise, the number of receptor clusters can be obtained from the experimentally measured noise σ_n^2. In particular, again accounting for the normalized output and using that $X_\mathrm{T} > R_\mathrm{T}$, we require that $\sigma_n^2 = \tilde{\alpha} N f^2 (1-f)^2 (1-p)^2 / R_\mathrm{T} = 0.092$ (see Eq. 4.37). Using $f = 0.5$, $p = 0.3$, the inferred $N = 12$, and $\tilde{\alpha} = 2$ [62], we then find an effective number of independent clusters of $R_\mathrm{T} \approx 8$. This estimate is low given the measured number of receptors, which is on the order of $10^3 - 10^4$ [49], but could be explained by very recent experiments, which show that the receptor array is tuned near a critical point, which effectively partitions the receptor array into a small number of large domains [35, 74]. In the main text we use $X_\mathrm{T} = 10^4$, which is in line with existing estimates [49], and entails that the phosphorylation noise is negligible.

5

PREDICTING CONCENTRATION CHANGES VIA DISCRETE RECEPTOR SAMPLING

To successfully navigate chemical gradients, microorganisms need to predict how the ligand concentration changes in space. Due to their limited size, they do not take a spatial derivative over their body length but rather a temporal derivative, comparing the current signal with that in the recent past, over the so-called adaptation time. This strategy is pervasive in biology, but it remains unclear what determines the accuracy of such measurements. Using a generalized version of the previously established sampling framework, we investigate how resource limitations and the statistics of the input signal set the optimal design of a well-characterized network that measures temporal concentration changes: the Escherichia coli chemotaxis network. Our results show how an optimal adaptation time arises from the trade-off between the sampling error, caused by the stochastic nature of the network, and the dynamical error, caused by uninformative fluctuations in the input. A larger resource availability reduces the sampling error, which allows for a smaller adaptation time, thereby simultaneously decreasing the dynamical error. Similarly, we find that the optimal adaptation time scales inversely with the gradient steepness, because steeper gradients lift the signal above the noise and reduce the sampling error. These findings shed light on the principles that govern the optimal design of the E. coli chemotaxis network specifically, and any system measuring temporal changes more broadly.

The contents of this chapter have been published in Physical Review Research **6**, 03049 (2024) [86].

Organisms ranging from bacteria to mammals have learned to navigate their environment in order to find food and avoid threats. Successful navigation requires the organism to predict the spatial structure of its surroundings, which necessitates measuring and storing relevant environmental properties. Therefore, how accurately these signals are sensed can fundamentally limit the success of navigation [20]. This in turn raises the question how accurately such signals can be transduced.

Microorganisms that navigate chemical gradients need to determine the correct direction to move in, which entails predicting the change in concentration that they will encounter, rather than its value. Because these organisms are typically small relative to the gradient length, the measurement error is large compared to the concentration difference over their body length [87]. Therefore, they cannot directly measure the gradient. Instead, these microorganisms only have access to the local concentration. Yet, they can also store past concentrations. How these cells should integrate the current and past information to predict the concentration change remains however unclear. In principle, cells can combine the concentration value with its derivative to predict the concentration change, and the optimal strategy for combining this information depends on the statistics of the environment. If the range of background concentrations is large compared to the typical concentration change over the signal correlation time as set by the organism's own motion, then the optimal system for predicting the concentration change is one that exhibits perfect adaptation (Chapter 4). It means that the organism bases its prediction on the concentration change only.

Interestingly, various organisms have indeed been shown to employ this strategy. A canonical example is the bacterial chemotaxis system, which is widely conserved across species [19, 21–24]. But also eukaryotic sperm cells measure temporal changes when navigating towards an egg [25–27], and even the multicellular nematode *Caenorhabditis elegans* depends on temporal derivatives in a range of taxis behaviors [28].

Even though measuring temporal changes appears to be a common and important function, it is not clear what sets the accuracy of such measurements. The fundamental information processing devices that allow living cells to measure concentration changes are biochemical signaling networks. Like any device, the accuracy of such networks is limited by the physical resources required to build and operate them, such as energy, components, and time. Here, we investigate how these resources limit the accuracy with which cells can predict changes in the encountered concentration during navigation. Specifically, we ask what determines the optimal design of the signaling network under limited resource availability?

To measure a temporal change, cells subtract from the most recent signal the signal further back into the past. The latter is performed via the adaptation system. Crucially, to yield a response of non-zero amplitude, which is necessary to lift the signal above the inevitable biochemical noise, the system cannot adapt instantly; it therefore cannot take an instantaneous derivative. On the other hand, the adaptation time should not be too long, because then the temporal derivative is taken over a larger window stretching further back into the past, which is less informative about the current or future derivative that the cell needs to predict. Therefore, there exists an optimal adaptation time that arises from this trade-off between a derivative that is most recent and one that is most reliable (Chapter 4). However, what precisely controls the optimal adaptation time, and how this depends on the statistics of the input and the available resources such as receptor and readout copies, remains unknown.

An intuitive perspective that is ideally suited to answer these questions is the previously established sampling framework [13, 15, 29]. This framework views the signaling network downstream of the receptor as a device that discretely samples the state of the receptor. From this starting point, it enables identification of the different contributions that comprise the full sensing error: the sampling error, caused by fluctuations in the number of samples, the binary nature of the receptor state, and receptor-level noise; and the dynamical error, resulting from uninformative fluctuations in the input. While previous work has used the sampling framework to investigate sensing the current signal, we here generalize and extend it to include the prediction of signal properties a specified time into the future. We then apply this generalized sampling framework to the *Escherichia coli* chemotaxis network: a well-characterized example of a network which measures temporal changes. We model the input signal after the experimentally measured input for *E.coli* chemotaxis in shallow gradients [20].

Our results distinctly show how an optimum for the adaptation time arises from its opposing effects on the sampling error and the dynamical error. While the former decreases with the adaptation time, the latter increases with it. Given the adaptation time, a larger number of receptor and readout molecules reduces the sampling error, shifting the balance between the sampling and dynamical error. Therefore, increasing the resource availability reduces the optimal adaptation time. Similarly, we find that the optimal adaptation time scales inversely with the steepness of the chemical gradient in which the organism navigates. The reason is that in a steeper gradient, the signal is more easily distinguished from the noise under the same resource availability. This again means that the sampling error decreases relative to the dynamical error, reducing the optimal adaptation time to decrease the latter. Finally, if the dynamics of the concentration change are Markovian, the optimal adaptation time is independent of the prediction interval. These findings likely extend well beyond *E. coli*, and have implications for the optimal design of any system that measures temporal changes, be it natural or man-made.

5.1. THEORY: SAMPLING FRAMEWORK

In general, the function of a biochemical signaling network is to estimate the value of a signal of interest, which typically varies in time. Sensing entails estimating the value of the signal at the current time t_0, while predicting the future state of the environment implies estimating the value a time τ into the future. To extend the sampling framework to be applicable to prediction as well as sensing we define the signal of interest as $s_\tau \equiv s(t_0+\tau)$ with $\tau \geq 0$. In this work we consider a time-varying input signal described by stationary Gaussian statistics (see Section 5.2).

In biochemical signaling networks, the activity state of receptor proteins is altered by ligand molecules that bind them. In turn, downstream readout proteins stochastically sample the receptor state $n \in \{0,1\}$. From these samples the signal of interest must then be inferred. A canonical motif that samples the activity state of upstream receptor proteins is the push-pull network (see also Chapter 3) [17]. In this network a sample of the

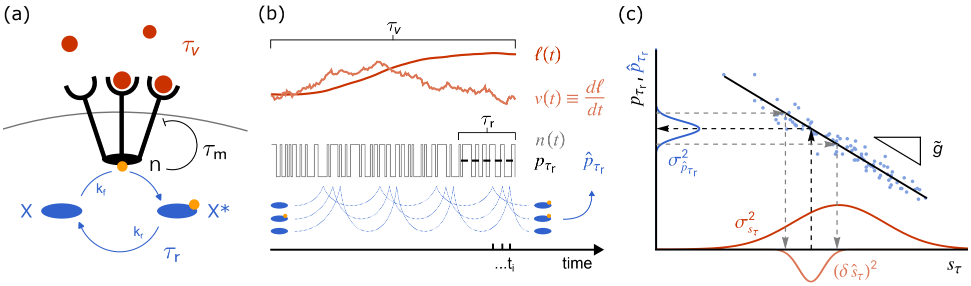


Figure 5.1: A push-pull motif samples the binary state of the chemotaxis receptor cluster. (a) Ligand binding affects the probability of a chemotaxis receptor cluster to reside in its active or inactive conformation. This binary cluster state n controls the methylation dynamics of its constituent receptors, leading to negative feedback on the adaptation timescale τ_m. The cluster state is sampled by the readout molecules X on the response timescale τ_r. (b) We consider an input signal defined by its concentration $\ell(t)$ and concentration derivative $v(t)$, with correlation time τ_v (Eqs. 5.12, 5.13 and 5.14). The instantaneous cluster activity $n \in \{0,1\}$ switches fast relative to the input correlation time, response time τ_r, and adaptation time τ_m. Due to the negative feedback, the mean cluster activity reflects the change in concentration over the past adaptation time τ_m. The network makes an estimate $\hat{p}_{\tau_r}$ of the cluster activity over the past response time τ_r by discretely sampling the instantaneous cluster state via the push-pull motif [panel (a)]; the estimate $\hat{p}_{\tau_r} = x^*/\overline{N}$ is given by the current number of active readout molecules x^*, reflecting the number of samples of active receptor clusters during the past integration time τ_r, over the mean number of samples $\overline{N}$ during this time τ_r (Eq. 5.5). (c) For linear Gaussian systems the future signal s_τ maps onto a current mean cluster activity over the response time p_{τ_r} via the dynamic input-output relation of Eq. 5.1. The variance in the estimate $\hat{p}_{\tau_r}$ given a signal value s_τ is the prediction error $\sigma^2_{\hat{p}_{\tau_\mathrm{r}}}$. Mapping the prediction error back onto the signal gives the network's error in the signal estimate $(\delta\hat{s}_\tau)^2$. The ratio between the total variance in the signal $\sigma^2_{s_\tau}$ and the error in the signal estimate $(\delta\hat{s}_\tau)^2$ is the signal-to-noise ratio (Eq. 5.2).

receptor state is stored in the chemical modification state of a readout protein, which decorrelates from the receptor state over the response time τ_r [Fig. 5.1(a)].

To estimate the signal value s_τ a time τ into the future, the cell integrates the receptor activity over a time τ_r, leading to an estimate $\hat{p}_{\tau_r}$ of the average receptor activity p_{τ_r} over the integration time τ_r [Fig. 5.1(b)]. However, during this past time τ_r, the input signal varies over its own timescale τ_v, which leads to changes in the receptor activity on this timescale as well [15, 29]. On top of variation on the timescale of the input dynamics, the receptor activity fluctuates on the timescale of ligand binding and unbinding, and on the timescale of the adaptation mechanism τ_m. In the linear regime, the dynamic input-output relation between the average receptor activity p_{τ_r} and the signal of interest s_τ is given by

$$p_{\tau_\mathrm{r}}(s_\tau) \equiv \mathbb{E}[\langle n(t_i)|s_\tau\rangle]_{t_i} = p + \tilde{g}s_\tau, \tag{5.1}$$

where the angle brackets denote an ensemble average over all receptors, $\mathbb{E}[\dots]_{t_i}$ is an average over all sampling times t_i, which are exponentially distributed over the integration time τ_r (Eq. 5.35), and $p \equiv \mathbb{E}[\langle n(t_i)\rangle]_{t_i}$ is the average receptor activity over all signal values. The dynamic input-output relation thus gives the average receptor activity p_{τ_r} over the response time τ_r given that the future signal is s_τ; p_{τ_r} is thus an average over all sources of noise, arising from receptor-ligand binding and receptor methylation, readout activation, and fluctuations in the past input that are not informative because they map onto the same future signal s_τ (see Fig. 5.2). The slope of the mapping between s_τ and p_{τ_r} is the dynamic gain $\tilde{g}$ [Fig. 5.1(c)] [37].

The accuracy of any signaling device can be quantified using the signal-to-noise ratio (SNR), which is a measure for the number of distinct signal values the system can distinguish. For systems with Gaussian statistics, as studied here, the SNR is given by the ratio of the signal variance $\sigma^2_{s_\tau}$ over the error in the cell's estimate of the signal $(\delta\hat{s}_\tau)^2 \equiv \mathbb{E}\left[\mathrm{Var}\left(\hat{s}_\tau|s_\tau\right)\right]_{s_\tau}$, i.e. the variance of the cell's signal estimate $\hat{s}_\tau$ under a fixed signal s_τ, averaged over all s_τ:

$$\mathrm{SNR} \equiv \frac{\sigma^2_{s_\tau}}{(\delta\hat{s}_\tau)^2} = \frac{\tilde{g}^2\sigma^2_{s_\tau}}{\sigma^2_{\hat{p}_{\tau_\mathrm{r}}}}. \tag{5.2}$$

The cell estimates the signal s_τ from the average receptor activity over the integration time, p_{τ_r}, via the dynamic input-output relation, see Eq. 5.1 and Fig. 5.1(c). Using the rules of error propagation, the error in the signal estimate is thus given by

$$(\delta\hat{s}_\tau)^2 = \sigma^2_{\hat{p}_{\tau_\mathrm{r}}}/\tilde{g}^2, \tag{5.3}$$

where the error in the estimate of the receptor activity $\hat{p}_{\tau_r}$ over the integration time τ_r is defined as

$$\sigma^2_{\hat{p}_{\tau_\mathrm{r}}} \equiv \mathbb{E}\left[\mathrm{Var}\left(\hat{p}_{\tau_r}|s_\tau\right)\right]_{s_\tau} \tag{5.4}$$

The signal to noise ratio of Eq. 5.2 also specifies the Gaussian mutual information between the signal and the network output [4].

To quantify the error in the cell's estimate of the receptor activity (Eq. 5.4), we have to consider how the cell makes this estimate. As a model system to investigate networks that measure changes in the input we use the *E. coli* chemotaxis network. In this network, the activity of a receptor cluster reflects the change in signal concentration over the past adaptation time τ_m (see Section 5.3 for details). Downstream of the cluster, its activity state is sampled via a push-pull motif [Fig. 5.1(a)] [17]. The cell's estimate of the fraction of active clusters is given by (also see [13])

$$\hat{p}_{\tau_r} = \frac{1}{N}\sum_{i=1}^{N} n_i(t_i) = \frac{x^*}{N}, \tag{5.5}$$

where $n_i(t_i) \in \{0,1\}$ is the outcome of sample i at sampling time t_i, which is set by the binary activity state of the receptor cluster that was sampled at time t_i [Fig. 5.1(b)]. The physical readout of the network is the number of active readout molecules $x^* = \sum_{i=1}^{N} n_i(t_i)$, which have been phosphorylated by an active receptor cluster. Since readout phosphorylation is driven by ATP hydrolysis, we consider the sampling process in the irreversible limit.

The number of samples N is set by the rate of sampling r and the timescale over which samples remain correlated with the receptor state, which is set by the integration, or response time τ_r. In the push-pull motif the sampling rate is set by the forward rate constant k_f, the number of receptor clusters R_T, and the number of available readout molecules X: $r = k_f x R_\mathrm{T}$ [Fig. 5.1(a)]. We assume that N is Poisson distributed with mean $\overline{N} = \bar{r}\tau_\mathrm{r}$. This mean number of samples can be expressed in terms of the steady state fraction of phosphorylated readouts $f = k_f p R_\mathrm{T}\tau_\mathrm{r}$ and the total number of readouts X_T [29],

$$\overline{N} = f(1-f)X_\mathrm{T}/p. \tag{5.6}$$

The steady state flux of readout molecules is given by $\bar{r}p = f(1-f)X_\mathrm{T}/\tau_\mathrm{r}$.

Using the definition of the cell's estimate of the receptor activity (Eq. 5.5) the error in this estimate (Eq. 5.4) can be decomposed into independent parts in a very general manner. We set out this decomposition in the section that follows. After the decomposition of the error we describe the dynamics and statistics of the input of the chemotaxis network (Section 5.2). Subsequently we introduce the chemotaxis network in more detail, and compute the dynamic gain $\tilde{g}$ (see Eq. 5.1), and the different contributions to the error in terms of the parameters of the system (Section 5.3). We then compute the full expression for the SNR and investigate its behavior as a function of the prediction interval, the resource availability, and the adaptation time (Sections 5.4-5.6). We compare the predictions of our theory to available experimental data on the *E. coli* chemotaxis network in Section 5.7. Finally, we illustrate the effect of the signal statistics on our results by computing the relative error for a different set of statistics (Section 5.8).

THE ERROR IN THE ESTIMATE OF THE RECEPTOR ACTIVITY

We can derive a general expression for the prediction error $\sigma^2_{\hat{p}_{\tau_\mathrm{r}}}$, which shows how the complete error decomposes into independent parts. We start from the definition of the error (Eq. 5.4), which we rewrite using the law of total variance

$$\begin{aligned}\sigma^2_{\hat{p}_{\tau_\mathrm{r}}} &= \mathrm{Var}\left(\hat{p}_{\tau_r}\right) - \mathrm{Var}\left(\mathbb{E}\left[\hat{p}_{\tau_r}|s_\tau\right]\right),\\ &= \mathrm{Var}\left(\mathbb{E}\left[\hat{p}_{\tau_r}|N\right]\right) + \mathbb{E}\left[\mathrm{Var}\left(\hat{p}_{\tau_r}|N\right)\right] - \mathrm{Var}\left(\mathbb{E}\left[\hat{p}_{\tau_r}|s_\tau\right]\right),\end{aligned} \tag{5.7}$$

where in the first line we use that the total variance in the estimate of the activity $\mathrm{Var}\left(\hat{p}_{\tau_r}\right)$, is the sum of the variance in the mean of $\hat{p}_{\tau_r}$ given s_τ, $\mathrm{Var}\left(\mathbb{E}\left[\hat{p}_{\tau_r}|s_\tau\right]\right)$, and the mean of the variance in $\hat{p}_{\tau_r}$ conditional on s_τ, $\mathbb{E}\left[\mathrm{Var}\left(\hat{p}_{\tau_r}|s_\tau\right)\right]$, which is the error $\sigma^2_{\hat{p}_{\tau_\mathrm{r}}}$ (Eq. 5.4). Indeed, the error in the estimate is its total variance minus the part which is informative about the signal of interest s_τ. Subsequently, in the second line, we split the total variance in the estimate $\hat{p}_{\tau_r}$ into a part that arises from fluctuations in the number of samples N, the first RHS term, and the mean variance in $\hat{p}_{\tau_r}$ when N is fixed, the second RHS term.

In Appendix 5.A we show how each term of Eq. 5.7 can be simplified further using the definition of the cell's estimate $\hat{p}_{\tau_r}$ (Eq. 5.5). The first term, the error caused by fluctuations in the number of samples, is given by

$$\mathrm{Var}\left(\mathbb{E}\left[\hat{p}_{\tau_r}|N\right]\right) = \frac{p^2}{\overline{N}}, \tag{5.8}$$

with the average cluster activity $p \equiv \mathbb{E}\left[\langle n(t_i)\rangle\right]_{t_i}$. As shown in previous work, this error would be zero if the sampled cluster functions bidirectionally, i.e. if inactive clusters would dephopshorylate readout molecules [13]; in contrast, in the chemotaxis network deactivation is not driven by inactive receptor clusters but rather by an enzyme (CheZ) independent of the receptor state, and then this term is non-zero. The fluctuations under a fixed number of samples, the second RHS term of Eq. 5.7, can be decomposed further into three parts:

$$\mathbb{E}\left[\mathrm{Var}\left(\hat{p}_{\tau_r}|N\right)\right] = \frac{p(1-p)}{\overline{N}} + \mathbb{E}\left[\mathrm{Cov}\left(n_i(t_i), n_j(t_j)|\boldsymbol{s}\right)\right]_{t_i,t_j,\boldsymbol{s}} + \mathrm{Var}\left(\mathbb{E}\left[\langle n(t_i)|\boldsymbol{s}\rangle\right]_{t_i}\right), \tag{5.9}$$

where the first part reflects the instantaneous variance of each sampled cluster, the second part is the cluster covariance under a fixed past signal trajectory $\boldsymbol{s} \equiv \{s(t)\}_{t\leq t_0}$, and the third part quantifies the effect of the signal history $\boldsymbol{s}$ on the activity of the cluster. Finally, the variance that is informative of the future signal value, i.e. the third RHS term of Eq. 5.7, is given by

$$\mathrm{Var}\left(\mathbb{E}\left[\hat{p}_{\tau_r}|s_\tau\right]\right) = \mathrm{Var}\left(\mathbb{E}\left[\langle n(t_i)|s_\tau\rangle\right]_{t_i}\right) = \tilde{g}^2\sigma^2_{s_\tau}, \tag{5.10}$$

which follows directly from the dynamic input output relation in Eq. 5.1. Substituting Eqs. 5.8, 5.9 and 5.10 into Eq. 5.7 yields the full prediction error

$$\sigma^2_{\hat{p}_{\tau_r}} = \underbrace{\frac{p^2}{\overline{N}} + \frac{p(1-p)}{\overline{N}} + \mathbb{E}\left[\mathrm{Cov}\left(n_i(t_i), n_j(t_j)|\boldsymbol{s}\right)\right]_{t_i,t_j,\boldsymbol{s}}}_{\text{sampling error}} + \underbrace{\mathrm{Var}\left(\mathbb{E}\left[\langle n(t_i)|\boldsymbol{s}\rangle\right]_{t_i}\right) - \tilde{g}^2\sigma^2_{s_\tau}}_{\text{dynamical error}}. \tag{5.11}$$

The first three terms together make up the sampling error. This error arises due to the stochastic nature of the sampling process downstream of the receptor, receptor-ligand binding and unbinding, and the adaptation mechanism. In this work we integrate out ligand binding, and we will therefore find that receptor methylation constitutes the only noise source on the receptor level. The sampling error quantifies all variability in the output under a constant input, as in [13] (see Fig. 5.2). The final two terms constitute the dynamical error; this is the error that arises from fluctuations in $\hat{p}_{\tau_r}$ that are caused by differences between past signal trajectories that map onto the same future signal of interest. These fluctuations contribute to the error in $\hat{p}_{\tau_r}$ because they do not provide any information on the future signal of interest [15] (Fig. 5.2).

Equation 5.11 reflects that, in general, the error for a linear sensing system can be decomposed into a contribution that arises from the stochastic sampling of the signal and a contribution that comes from the fact that not all signal fluctuations in the past correspond to the signal which the cell aims to predict. More specifically, Eq. 5.11 holds for any cellular sensing system in which the signal is inferred from the receptor activity, estimated using the downstream signaling system as a sampling device, as in Eq. 5.5. Yet, to derive the sensing error SNR^{-1} (see Eqs. 5.2, 5.3 and 5.4) for the chemotaxis network, we need to evaluate the sampling error and the dynamical error, as well as the dynamic gain $\tilde{g}$ (see Eq. 5.3). These quantities depend on the specific characteristics of the sensing system and the signal statistics, discussed next.

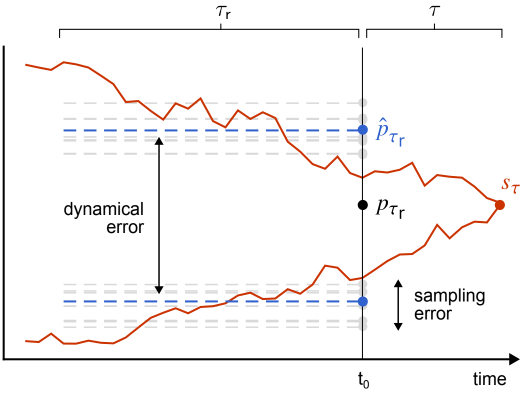


Figure 5.2: The total error in the cell's estimate of the receptor activity can be decomposed into the dynamical error and the sampling error. For linear signaling systems, a given current or future signal s_τ (red dot) maps onto a single mean receptor activity p_{τ_r} at the current time t_0 (black dot) via the dynamic input-output relation of Eq. 5.1 [Fig. 5.1(c)]. However, the past input and thus receptor activity on which the estimate $\hat{p}_{\tau_r}$ (blue dot) is based varies in time, leading to a dynamical error. This error arises because different past trajectories of the signal map onto a common future value s_τ, leading to uninformative variations in $\hat{p}_{\tau_r}$. Even for a given input trajectory the receptor noise, which in this work is only caused by receptor methylation, and the stochastic nature of the sampling process downstream of the receptor, lead to deviations in the estimate $\hat{p}_{\tau_r}$ (gray dots) which constitute the sampling error.

5.2. SIGNAL STATISTICS

In general, it is hard to know what the natural input statistics are that an organism experiences, and these may vary widely. We can start from the observation that microorganisms in dilute environments are faced with chemical gradients that are exceedingly shallow compared to their own length. In such environments, the only signal property that the cell can measure is the local concentration. But to determine if it is moving in the right direction, the cell must predict the change in concentration over time. So, while the cell can only measure concentrations, it is interested in the concentration's temporal derivative.

An ideal model system to study networks that can predict temporal changes is the *E. coli* chemotaxis network, as we also studied in Chapter 4. *E. coli* swims in its environment with a speed which exhibits persistence. This leads to an auto-correlation function for the concentration change which does not decay instantaneously [20]. To model a signal which is characterized by both the concentration and its derivative, and in which correlations in the derivative persist over the correlation time set by the motion of the cell, we again use the classical model of a particle in a harmonic well [38], as studied in Chapters 2 and 4,

$$\delta\dot{\ell} = v(t), \tag{5.12}$$

$$\dot{v} = -\omega_0^2 \delta\ell(t) - v(t)/\tau_v + \eta_v(t). \tag{5.13}$$

Here, $\delta\ell(t) \equiv (c(t) - c_0)/c_0$ is the relative deviation of the concentration $c(t)$ from its background value c_0. Note that this is a slight deviation from the model in Chapters 2 and 4, where we did not divide by c_0 (or equivalently $\bar{\ell}$). The derivative of this relative concentration is $v(t)$, and $\eta_v(t)$ is a Gaussian white noise process that drives the stochastic fluctuations in the signal. The parameter ω_0 sets the variance in the concentration σ_ℓ^2 relative to that in its derivative $\sigma_\ell^2 = \sigma_v^2/\omega_0^2$, where the variance in the derivative σ_v^2 is set by the swimming behavior of the cell. The relaxation time τ_v is set by the run duration, as this is the timescale over which the input fluctuations decorrelate.

Let us shortly recapitulate how, in Section 4.1, we obtained the relevant signal regime for *E. coli* performing chemotaxis in shallow gradients. The range of ligand concentrations which *E. coli* might encounter is very large, based on the dissociation constants of the inactive and active receptor conformations. For the Tar-MeAsp receptor ligand combination these respectively are $K_\mathrm{D}^\mathrm{I} = 18\mu\mathrm{M}$ and $K_\mathrm{D}^\mathrm{A} = 2900\mu\mathrm{M}$ [56–58]. This suggests that the total variance in the ligand concentration is much larger than the concentration change over the course of a run, i.e. $\sigma_\ell \gg \tau_v \sigma_v$ and thus $\omega_0 \ll \tau_v^{-1}$. In this regime, the correlation function of $v(t)$ becomes a simple exponential with variance σ_v^2 and decay time τ_v:

$$\langle \delta v(t) \delta v(t') \rangle = \sigma_v^2 e^{-|t-t'|/\tau_v}. \tag{5.14}$$

The correlation function of Eq. 5.14 corresponds to what has been observed experimentally for *E. coli* cells swimming in shallow exponential concentration gradients [20]. When cells swim in shallow gradients, i.e. with a characteristic length much longer than the length of a run, they swim as if there is no gradient. The correlation function of the positional velocity $v_x(t)$ in the absence of a gradient has been measured to be an exponential with variance $\sigma_{v_x}^2$ and decay time τ_v set by the duration of a run [20]. This can be mapped onto the correlation function of Eq. 5.14, where $v(t) \equiv c_0^{-1} dc/dt$, when we consider that the concentration gradient is given by $c(t) = c_0 \exp[gx(t)]$ with the gradient steepness g. We find for the absolute concentration change over time $dc/dt = dc/dx\, dx/dt = gc(t)v_x(t)$, and thus we have for variance of the relative concentration change $v(t)$:

$$\sigma_v^2 = g^2 \sigma_{v_x}^2. \tag{5.15}$$

Experimental measurements provide the relaxation time $\tau_v^{-1} = 0.86\mathrm{s}^{-1}$ and the variance of the positional derivative $\sigma_{v_x}^2 = 157\mu\mathrm{m}^2\mathrm{s}^{-2}$ [20].

5.3. CHEMOTAXIS MODEL

Here we shortly summarize our model of the chemotaxis network, as discussed in detail in Section 4.2. We reiterate some of the most important aspects of the model, and we explain its central properties that are important for the results presented in this chapter. A derivation of these properties based on the sampling framework discussed above, is presented in Appendix 5.B.

In the *E. coli* chemotaxis network, receptors cooperatively control the activity of the kinase CheA, which controls the phosphorylation of the readout protein CheY [Fig. 5.1(a)] [60–63]. The receptor cooperativity has been successfully described using the Monod-Wyman-Changeux (MWC) model, where individual receptors are assumed to form clusters in which all receptors must reside in the same activity state [20, 50, 52, 58, 62, 65–67].

Furthermore, inactive receptors are methylated by the enzyme CheR, which increases the probability for the cluster to be active, and active receptors are demethylated by CheB. These methylation dynamics ensure that the network exhibits perfect adaptation with respect to the background concentration [19, 57, 64, 88–90]. Therefore, the activity state of the cluster only responds transiently to changes in the input, and reflects the recent change in concentration.

Because both ligand binding and switching between the active and inactive state of the cluster are fast compared to the input, methylation, and phosphorylation dynamics, it is instructive to take a quasi-equilibrium approach and consider the average cluster activity given the methylation level of the cluster and the extracellular ligand concentration. In the linear noise approximation we have for the activity (see Appendix 5.B)

$$a(t) \equiv \langle n(t) | \delta m, \delta \ell \rangle = p + \alpha \delta m(t) - \beta \delta \ell(t), \tag{5.16}$$

where p is the mean activity, $\delta m(t)$ represents the methylation level of the cluster, and $\delta \ell(t)$ represents the ligand concentration, both defined as deviations from their mean. The constants α and β respectively depend on the free energy cost of methylation $\tilde{\alpha}$, and on the dissociation constants $K_{\mathrm{D}}^{\mathrm{I}}$ and $K_{\mathrm{D}}^{\mathrm{A}}$ and background concentration c_0. The methylation dynamics are given by,

$$\delta \dot{m} = -\delta a(t)/(\alpha \tau_{\mathrm{m}}) + \eta_m(t), \tag{5.17}$$

where τ_{m} is the adaptation time, and η_m is Gaussian white noise (see Eq. 5.51).

The dynamic gain of the network maps the signal of interest onto the receptor activity [Eq. 5.1, Fig. 5.1(c)]. For the purpose of navigation, we define the signal of interest to be the change in concentration $v_\tau \equiv v(t_0 + \tau)$ some time $\tau \geq 0$ into the future. The autocorrelation of the change in concentration is given by Eq. 5.14. The dynamic gain of the chemotaxis network with respect to this signal of interest is (Eqs. 5.54-5.57),

$$\tilde{g} = \frac{g_{v \to p} e^{-\tau/\tau_v}}{(1 + \tau_{\mathrm{m}}/\tau_v)(1 + \tau_{\mathrm{r}}/\tau_v)} = \frac{-\tau_{\mathrm{m}} \beta e^{-\tau/\tau_v}}{(1 + \tau_{\mathrm{m}}/\tau_v)(1 + \tau_{\mathrm{r}}/\tau_v)}, \tag{5.18}$$

where τ_v is the signal correlation time, τ_{r} is the network response time, τ_{m} is the adaptation time, and the static gain from the input signal derivative v to the steady state activity p is given by

$$g_{v \to p} \equiv \partial_v p = -\tau_{\mathrm{m}} \beta. \tag{5.19}$$

Equation 5.18 shows that the dynamic gain $\tilde{g}$ is maximized for a fast response $\tau_{\mathrm{r}} \ll \tau_v$, and slow adaptation $\tau_{\mathrm{m}} \gg \tau_v$. A longer adaptation time increases the dynamic gain via the static gain (Eq. 5.19), because the absolute difference between sequential inputs is on average larger over this longer time. Yet, the dynamic gain saturates as τ_{m} increases:

$$\lim_{\tau_{\mathrm{m}} \to \infty} \tilde{g} = \frac{-\tau_v \beta e^{-\tau/\tau_v}}{1 + \tau_{\mathrm{r}}/\tau_v}. \tag{5.20}$$

In this limit, considering that typically $\tau \leq \tau_v$ and $\tau_{\mathrm{r}} \ll \tau_v$, the dynamic gain is approximately proportional to the signal correlation time τ_v. The reason is that fluctuations further than τ_v in the past cannot affect the mapping from the current signal, which is most correlated to the signal of interest v_τ, to the current receptor state. Finally, increasing the prediction interval τ reduces the dynamic gain because the correlation between future signal and sensed input decreases.

To determine the sampling error of the chemotaxis network, we require the cluster covariance under a fixed input signal, which is the third RHS term in Eq. 5.11. In our chemotaxis model, this covariance is a consequence of the methylation noise only and it is given by (Eqs. 5.58-5.63)

$$\mathbb{E}\left[\mathrm{Cov}\big(n_i(t_i), n_j(t_j)|\boldsymbol{s}\big)\right]_{t_i,t_j,\boldsymbol{s}} = \frac{\alpha p(1-p)}{R_\mathrm{T}(1+\tau_\mathrm{r}/\tau_\mathrm{m})}, \approx \alpha p(1-p)/R_\mathrm{T}. \tag{5.21}$$

Here, p is the mean cluster activity and R_T is the total number of independent receptor clusters. Substitution of Eq. 5.21 in Eq. 5.11 yields the full sampling error of the chemotaxis network

$$\sigma^{2,\mathrm{samp}}_{\hat{p}_{\tau_\mathrm{r}}} = \frac{p^2}{\overline{N}} + \frac{p(1-p)}{\overline{N}_\mathrm{I}}, \tag{5.22}$$

with the number of independent samples

$$\overline{N}_\mathrm{I} \equiv f_\mathrm{I}\overline{N} = \frac{\overline{N}}{1+\overline{N}/R_\mathrm{I}}, \tag{5.23}$$

where $f_\mathrm{I} = 1/(1+\overline{N}/R_\mathrm{I})$ is the fraction of independent samples and

$$R_\mathrm{I} = R_\mathrm{T}(1+\tau_\mathrm{r}/\tau_\mathrm{m})/\alpha \tag{5.24}$$

the number of independent receptor states during an integration time τ_r.

In contrast to previous work [15, 29], the sampling error does not depend on the correlation time τ_c of receptor-ligand binding because here we have assumed that ligand binding is much faster than the response time τ_r. Still, the cluster state remains correlated over time due to receptor methylation and this means that the expression for the sampling error of the chemotaxis network studied here is almost identical to that of the push-pull network studied in [15, 29]. However, unlike ligand binding noise, the methylation noise cannot be averaged out because the methylation timescale τ_m is longer than the response time τ_r, i.e. $1+\tau_\mathrm{r}/\tau_\mathrm{m} \approx 1$ in Eq. 5.21. Moreover, because the methylation noise affects the receptor activity via the factor α (Eqs. 5.16 and 5.17), which controls how strongly methylation changes the receptor activity, the cluster covariance also increases with α, since it increases the temporal covariance within each cluster.

Equation 5.23 reflects that the number of receptor samples $\overline{N}$, which is proportional to the number of readouts X_T (Eq. 5.6), and the number of independent receptor states R_I, proportional to the number of receptor clusters R_T (Eq. 5.24), are fundamental resources that limit the sensing accuracy like weak links in a chain [13]: when $\overline{N} \gg R_\mathrm{I}$ the number of independent samples is limited by the number of receptor states and $\overline{N}_\mathrm{I} \approx R_\mathrm{I}$, and vice versa, when $R_\mathrm{I} \gg \overline{N}$ the total number of samples is limiting and $\overline{N}_\mathrm{I} \approx \overline{N}$ (also see Fig. 5.6 and Appendix 5.C).

Finally, to compute the dynamical error, we derive the variation in the network output that is caused by the past input trajectory (Eqs. 5.64-5.69):

$$\mathrm{Var}\left(\mathbb{E}\left[\langle n(t_i)|\boldsymbol{s}\rangle\right]_{t_i}\right) = \frac{g^2_{v\to p}\sigma^2_v}{(1+\tau_\mathrm{m}/\tau_v)(1+\tau_\mathrm{r}/\tau_v)}\left(1+\frac{\tau_\mathrm{m}\tau_\mathrm{r}}{\tau_v(\tau_\mathrm{m}+\tau_\mathrm{r})}\right), \tag{5.25}$$

with the static gain $g_{v\to p}$ given by Eq. 5.19. Just like the dynamic gain (Eq. 5.18) this variation is maximized for a fast response $\tau_\mathrm{r} \ll \tau_v$ and slow adaptation $\tau_\mathrm{m} \gg \tau_v$. Indeed, in the regime that $\tau_\mathrm{m} \gg \tau_v$ we have $\mathrm{Var}\left(\mathbb{E}\left[\langle n(t_i)|\boldsymbol{s}\rangle\right]_{t_i}\right) \propto \tau_\mathrm{m}$. Therefore, unlike the dynamic gain, Eq. 5.25 does not saturate for an increasing adaptation time. The reason is that more and more values of the historical input contribute to the variance in the output as long as the system does not adapt. Clearly, not all of the variation quantified by Eq. 5.25 will carry information about the signal of interest v_τ. Substituting Eq. 5.25 in Eq. 5.11 yields the the dynamical error, which is the total uninformative variation caused by the past input trajectory

$$\sigma^{2,\mathrm{dyn}}_{\hat{p}_{\tau_\mathrm{r}}} = \tilde{g}^2\sigma^2_v\left[e^{2\tau/\tau_v}\left(1+\frac{\tau_\mathrm{m}}{\tau_v}\right)\left(1+\frac{\tau_\mathrm{r}}{\tau_v}\right)\left(1+\frac{\tau_\mathrm{m}\tau_\mathrm{r}}{\tau_v(\tau_\mathrm{m}+\tau_\mathrm{r})}\right)-1\right], \tag{5.26}$$

with the dynamic gain $\tilde{g}$ of Eq. 5.18. Even though Eq. 5.26 is the dynamical error in predicting the concentration change (rather than the value), its form is strikingly similar to the dynamical error of a push-pull network that predicts the current concentration, derived in [15, 29]. The reason for this similar form is that both the concentration considered in these previous works and the concentration derivative considered here are Markovian signal properties. Indeed, if we consider a non-Markovian signal concentration and derivative, the dynamic gain and dynamical error change (see Section 5.8).

5.4. RELATIVE PREDICTION ERROR

The central result of this work is the relative error, i.e. SNR^{-1}, made by the *E. coli* chemotaxis network when it predicts the future concentration change. Using the definition of the signal-to-noise ratio (Eq. 5.2), with the dynamic gain given in Eq. 5.18, and the prediction error $\sigma^2_{\hat{p}_{\tau_\mathrm{r}}} = \sigma^{2,\mathrm{samp}}_{\hat{p}_{\tau_\mathrm{r}}} + \sigma^{2,\mathrm{dyn}}_{\hat{p}_{\tau_\mathrm{r}}}$ given by Eqs. 5.22 and 5.26, we obtain

$$\mathrm{SNR}^{-1} = \underbrace{\frac{e^{2\tau/\tau_v}}{\tau^2_\mathrm{m}\beta^2\sigma^2_v}\left(1+\frac{\tau_\mathrm{m}}{\tau_v}\right)^2\left(1+\frac{\tau_\mathrm{r}}{\tau_v}\right)^2\left(\frac{p^2}{\overline{N}}+\frac{p(1-p)}{\overline{N}_\mathrm{I}}\right)}_{\text{sampling error}} + \underbrace{e^{2\tau/\tau_v}\left(1+\frac{\tau_\mathrm{m}}{\tau_v}\right)\left(1+\frac{\tau_\mathrm{r}}{\tau_v}\right)\left(1+\frac{\tau_\mathrm{m}\tau_\mathrm{r}}{\tau_v(\tau_\mathrm{m}+\tau_\mathrm{r})}\right)-1}_{\text{dynamical error}}. \tag{5.27}$$

This expression is similar in structure to the relative error of the push-pull network without adaptation, which was derived in earlier work (Eq. 6 of [15]). The reason is that, while the adaptation system affects the receptor dynamics, the downstream push-pull motif still acts as a device that discretely samples the receptor state. As a result, the relative error has two contributions: the sampling error, which arises from the stochasticity

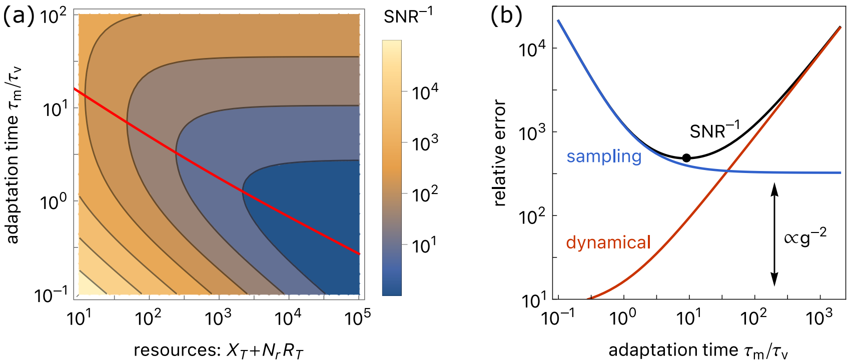


Figure 5.3: The relative error is set by the resource availability, adaptation time, and gradient steepness. (a) The relative error (SNR^{-1}, Eq. 5.27) as a function of the resource availability $C = X_T + N_r R_T$ and the adaptation time τ_m/τ_v. The relative error decreases monotonically with higher resource availability. The error is minimized for the optimal adaptation times indicated by the red line, which decreases with the resource availability. The ratio of readouts to receptors X_T/R_T obeys Eq. 5.29. (b) The relative dynamical error, sampling error, and their sum, the total relative error (SNR^{-1}, Eq. 5.27), as a function of the adaptation time τ_m/τ_v. The optimal adaptation time arises from a trade-off between the sampling error, which decreases with the adaptation time, and the dynamical error, which increases with the adaptation time. The minimal total error (black dot) occurs close to the point where the sampling error saturates as a function of τ_m/τ_v. The minimal sampling error is proportional to $1/g^2$. In (a) and (b) $g = 2\mathrm{mm}^{-1}$, in (b) $X_T = 10^4$ and $R_T = 8$ [20, 35, 38]. Other parameters: $N_r = 12$, $p = 0.3$, $f = f^{\mathrm{opt}} = 0.5$, $\tau_r = 0.1\mathrm{s}$, $\tau = \tau_v = 1.16\mathrm{s}$, $c_0 = 100\mu\mathrm{m}$, $\sigma_{v_x} = 157\mu\mathrm{m}^2\mathrm{s}^{-2}$ [20, 35, 38, 62]; $\tilde{\alpha} = 2k_BT$ [62]; $K_D^I = 18\mu\mathrm{M}$ and $K_D^A = 2900\mu\mathrm{M}$ [56–58]. Code to reproduce this figure is available at [91].

in sampling the state of the receptor, and the dynamical error, which arises from the dynamics of the input signal (see Fig. 5.2). However, while this expression for the relative error has a form that is similar to that for the push-pull network, there are also key differences.

First of all, both the sampling and the dynamical error depend on the forecast interval. In general, the dynamical error arises because while the system aims to predict the current or future derivative, it measures the change in concentration over the timescale τ_m on the level of the receptor, and reads out the receptor activity over the timescale τ_r (Fig. 5.2). The network thus only measures an instantaneous concentration change when both τ_m and τ_r go to zero. Still, even in this limit, the dynamical error remains finite as long as the forecast interval is larger than zero, due to the inherent unpredictability of the future signal.

Perhaps surprisingly, the relative sampling error also depends on the forecast interval τ. While the absolute sampling error of the network is independent of the forecast interval (Eq. 5.22), the dynamic gain does depend on it (Eq. 5.18). When the forecast interval increases, the dynamic gain decreases, reducing the effect of the signal of interest on the receptor activity. Therefore, while the absolute sampling error remains constant, the relative sampling error increases with the forecast interval. In short, for a larger forecast interval it becomes harder to lift the signal above the sampling noise.

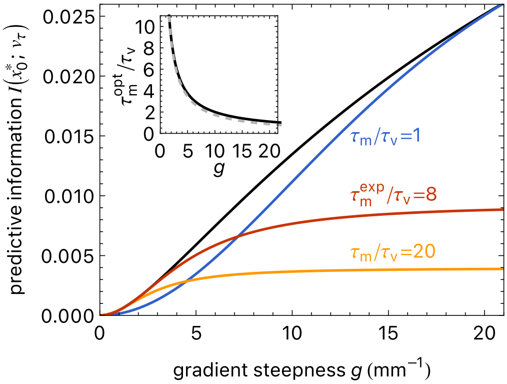


Figure 5.4: The optimal adaptation time depends on the gradient steepness. The predictive information $I(x_0^*; v_\tau) = I(\hat{p}_{\tau_r}; v_\tau) = 0.5\log(1+\text{SNR})$, with the SNR of Eq. 5.27, between the current number of phosphorylated readouts $x_0^* = \overline{N}\hat{p}_{\tau_r}$ (Eq. 5.5) and the future input derivative v_τ, for various adaptation times τ_m. Along the black curve, the adaptation time has been optimized; $\tau_\text{m}^\text{opt}/\tau_v$ as a function of the gradient steepness is shown in the inset. Experiments show that for *E. coli* the adaptation time is $\tau_\text{m}^\text{exp}/\tau_v \approx 8$ [19, 20, 62], which is close to optimal for $g \lesssim 4\text{mm}^{-1}$ (red curve). Reducing the adaptation time reduces the accuracy in shallow gradients and increases it in steeper gradients (blue curve), while increasing the adaptation time reduces the accuracy in steeper gradients but does not markedly increase the accuracy in shallow gradients (yellow curve). This suggests that the system has been been optimized for sensing shallow gradients. Inset: the optimal adaptation time $\tau_\text{m}^\text{opt}/\tau_v$ scales inversely with the gradient steepness g, numerical result (solid black line), and analytical approximation (dashed gray line, Eq. 5.31). Parameters are as in Fig. 5.3 and code to reproduce the figure is available at [91].

The second notable difference with the result on the push-pull network concerns the role of adaptation. It reflects the fact that the chemotaxis system takes a temporal derivative at the receptor level on a timescale set by the adaptation time. The dynamical error increases monotonically with the adaptation time τ_m, because for a longer adaptation time the system compares the current concentration to concentrations further in the past. Consequently, this change in concentration is less informative about the current derivative, which is the signal property most correlated to the future derivative (Eq. 5.14). In fact, when $\tau_\text{m} \to \infty$, the system does not adapt and the chemotaxis network therefore reduces to a push-pull network, which does not measure the derivative but rather the signal value. In this limit, the dynamical error diverges because the signal value is not correlated with the future derivative that the cell aims to predict. In contrast, the sampling error decreases monotonically with τ_m, because a longer adaptation time increases the dynamic gain (Eq. 5.18). How the optimal adaptation time that arises from these antagonistic effects depends on other parameters, such as the gradient steepness and the resource availability, is discussed in Section 5.6.

A third difference resides in the number of independent samples $\overline{N}_\text{I}$ (Eq. 5.23). For a push-pull network driven by a simple receptor, the number of independent samples is given by $\overline{N}_\text{I}^\text{PPN} = f_\text{I}^\text{PPN}\overline{N}_\text{eff}$, where the number of effective samples $\overline{N}_\text{eff} = \overline{N}$ in the irreversible limit, as we also study here [13, 15, 29]. For the push-pull network the fraction of independent samples can be expressed as $f_{\rm I}^{\rm PPN} = 1/(1+\overline{N}/R_{\rm I}^{\rm PPN})$ with the number of independent receptor states during an integration time $R_{\rm I}^{\rm PPN} = R_{\rm T}(1+\tau_{\rm r}/\tau_{\rm c})$, where $\tau_{\rm c}$ is the correlation time of the receptor binding state [13, 15, 29]. In our treatment of the chemotaxis model we consider the limit where $\tau_{\rm c} \ll \tau_{\rm r}$, in which case $R_{\rm I}^{\rm PPN}$ diverges and $f_{\rm I}^{\rm PPN} \approx 1$. However, the fraction of independent samples does not become unity for the chemotaxis network because the receptor state remains correlated due to the slow methylation dynamics, $\tau_{\rm m} \gg \tau_{\rm r}$. Therefore, the number of independent receptor states becomes limited by the number of receptor clusters and their covariance $R_{\rm I} \approx R_{\rm T}/\alpha$ (Eq. 5.23).

The sampling error can be mitigated in a number of ways. One is to increase the number of receptors per cluster $N_{\rm r}$, because this increases the magnitude of the static gain (see Eq. 5.19 where $\beta \propto N_{\rm r}$). Another is to simultaneously increase the total number of samples $\overline{N}$ and the number of independent samples $\overline{N}_{\rm I}$, which requires increasing both the number of readout molecules $X_{\rm T}$ and the number of receptor clusters $R_{\rm T}$ (Eqs. 5.6 and 5.23). Indeed, as observed for the push-pull network in earlier work [13, 15], these resources limit sensing and hence prediction like weak links in a chain (also see Fig. 5.6) and Appendix 5.C. However, increasing the cluster size, the number of clusters, or the number of readout molecules all require a larger number of proteins to be used by the network, which are resources that come at a physical cost.

5.5. Optimal resource allocation

To investigate how resources should be optimally allocated to minimize the sampling error (Eq. 5.27) we define a simple cost function, as in Chapter 4:

$$C = X_{\rm T} + N_{\rm r} R_{\rm T}, \tag{5.28}$$

where $X_{\rm T}$ is the number of readout molecules, $R_{\rm T}$ is the number of independent receptor clusters, and $N_{\rm r}$ is the number of receptors per cluster. This cost function captures the idea that a cell must choose whether it spends its resources on making more readout molecules on the one hand, or more receptors on the other. In general, the prediction error (Eq. 5.27) depends not only on the number of receptors and readout proteins, but also on the energy to drive the network [13, 15, 38]. These two constraints can be treated on the same footing by recognizing that the energy to synthesize the required proteins scales linearly with their copy number, with a proportionality constant set by the protein length. Here, we focus on the bound on the predictive information as set by the number of proteins, also because the effects of the energetic cost of driving the phosphorylation and methylation cycles on the optimal prediction strategy are relatively minor (Fig. 4.4). While the precise functional form of the constraint is somewhat arbitrary, the linear form of Eq. 5.28 is natural because the prediction error depends on the number of receptor and readout molecules, not on higher powers thereof.

Given a total resource availability C and a fixed number of receptors per cluster, the cell can tune the ratio of receptors to readouts. To determine what the optimal ratio is that minimizes the sampling error, we express both $R_{\rm T}$ and $X_{\rm T}$ in terms of their ratio and the total resource availability C, and we use that we can express the mean number of samples as in Eq. 5.6. Subsequently taking the derivative of Eq. 5.27 with respect to $X_\mathrm{T}/R_\mathrm{T}$ and equating to zero then gives the optimal ratio,

$$\left(\frac{X_\mathrm{T}}{R_\mathrm{T}}\right)^\mathrm{opt} = \frac{\sigma_X}{\sigma_R}\frac{p\sqrt{1+\tau_\mathrm{r}/\tau_\mathrm{m}}}{f(1-f)} \approx \frac{\sigma_X}{\sigma_R}\frac{p}{f(1-f)}, \tag{5.29}$$

where we have used that the adaptation time must be larger than the response time and thus $\sqrt{1+\tau_\mathrm{r}/\tau_\mathrm{m}} \approx 1$. We have further defined the noise per receptor $\sigma_R^2 \equiv \alpha p(1-p)/N_\mathrm{r} = \tilde{\alpha} p^2(1-p)^2$ (also see Eqs. 5.47, 5.49 and 5.63), and the noise per readout molecule $\sigma_X^2 \equiv f(1-f)$. In terms of $\overline{N}$, using Eq. 5.6, we find that Eq. 5.29 yields an intuitive relation for optimal networks,

$$\overline{N} = \frac{\sigma_X}{\sigma_R} R_\mathrm{T}. \tag{5.30}$$

This relation shows that for equal noise magnitudes per protein, the average number of samples should equal the total number of receptor clusters. This simple relation arises from the fact that the methylation noise cannot be averaged out, and a minimally redundant design is therefore one in which each receptor cluster is sampled once.

Given the optimal ratio of readouts to receptors in Eq. 5.29 we can compute the relative error (Eq. 5.27) as a function of the total resource availability C and the adaptation time τ_m [Fig. 5.3(a)]. As expected, we find that the error decreases monotonically with the resource availability. More interesting is that we find a clear optimum for the adaptation time τ_m.

5.6. OPTIMAL ADAPTATION TIME

The optimal adaptation time τ_m, given by the red line in Fig. 5.3(a), arises from the antagonistic effect of the adaptation time on the sampling error and the dynamical error [Fig. 5.3(b)]. The sampling error decreases monotonically with the adaptation time because a longer adaptation time increases the change in the receptor activity upon the same change in the current or future signal derivative, i.e. it increases the (dynamic) gain (Eqs. 5.18, 5.19 and 5.27). However, increasing the adaptation time means that the derivative is taken over a longer time further back into the past, and this derivative will be less informative about the future derivative that the cell aims to predict: the dynamical error increases monotonically with τ_m (Eq. 5.27). The minimal total error occurs for the smallest adaptation time that is sufficiently large to lift the signal above the noise, i.e. reduce the sampling error, while minimizing the dynamical error [Fig. 5.3(b)].

The value of the adaptation time for which the total error is minimized depends on the resource availability C and the gradient steepness g: these parameters set the magnitude of the sampling error (Eqs. 5.15 and 5.27). To obtain analytical insight into the optimal adaptation time $\tau_\mathrm{m}^\mathrm{opt}$, we exploit that the response time τ_r must be smaller than the adaptation time τ_m to mount a non-zero response to transient input changes. We further consider that the relevant regime for *E. coli* is likely that where gradients are shallow relative to the length of a run (also see Section 5.7). This means that the sampling error dominates over the dynamical error, although the latter is not negligible (see Eq. 5.27 with $\sigma_v^2 = g^2\sigma_{v_x}^2$, Eq. 5.15). To minimize the prediction error (Eq. 5.27) in this regime, the adaptation time must be large relative to the signal correlation time $\tau_{\mathrm{m}} \gg \tau_v$, which is set by the duration of a run. We obtain for the optimal adaptation time (see Appendix 5.C)

$$\tau_{\mathrm{m}}^{\mathrm{opt}} \approx \frac{\sqrt{2}}{\beta g \sigma_{v_x}} \sqrt{\frac{p^2}{\overline{N}} + \frac{p(1-p)}{\overline{N}_{\mathrm{I}}}}, \quad \text{for } \tau_{\mathrm{m}} \gg \tau_v, \tau_{\mathrm{r}}, \tag{5.31}$$

where the number of independent samples $\overline{N}_{\mathrm{I}}$ is given by Eq. 5.23 with $R_{\mathrm{I}} = R_{\mathrm{T}}/\alpha$. The inset of Fig. 5.4 shows that Eq. 5.31 is a good approximation of the optimal adaptation time over a large range of the gradient steepness g.

Equation 5.31 shows how the optimal adaptation time decreases when we increase the total number of receptor samples $\overline{N} \propto X_{\mathrm{T}}$ (Eq. 5.6) or the number of independent receptor samples $\overline{N}_{\mathrm{I}}$, which depends on both X_{T} and R_{T} (Eq. 5.23). This is because increasing these resources reduces the sampling error: decreasing the adaptation time then decreases the dynamical error more than it increases the relatively small sampling error. In Appendix 5.C, we discuss in more detail how the optimal adaptation time varies with X_{T} and R_{T} separately. Equation 5.31 also shows that the optimal adaptation time decreases as the gradient steepness increases. The reason is that a steeper gradient generates a stronger signal with a larger variance (Eq. 5.15), which reduces the sampling error (Eq. 5.27).

5.7. Comparison to experiment

To check whether the uncovered design principles (Eqs. 5.29 and 5.31) are relevant to real world biochemical networks, we evaluate the design of the *E. coli* chemotaxis network in this light.

To assess the design principle of Eq. 5.29, we use the definitions of σ_X and σ_R given below it. For p and f of order 1/2 and $\tilde{\alpha} = 2$, based on experiment [62], Eq. 5.29 predicts an optimal number of readout molecules per receptor cluster of $X_{\mathrm{T}}/R_{\mathrm{T}} \approx 3$. This is in good agreement with earlier predictions [13] and the experimental data of Li and Hazelbauer [49], assuming a cluster consists of 2 trimers of receptor dimers and 2 CheA dimers [92]. With $X_{\mathrm{T}} \sim 10^3 - 10^4$ readout molecules depending on the growth rate [49], this result, i.e. $X_{\mathrm{T}}/R_{\mathrm{T}} \approx 3$, suggests that the number of receptor clusters is in the range $R_{\mathrm{T}} \sim 10^2 - 10^3$. On the other hand, fitting more recent experimental data with an MWC based chemotaxis model as we use here, suggests a much smaller number of receptor clusters of $R_{\mathrm{T}} \approx 8$ (Section 4.5 and Appendix 4.C). However, this estimate for the number of receptor clusters was based on fitting the noise amplitude of the model to the experimental data of [20]. Recent experiments indicate that the receptor array is poised to a critical point [74], where receptor switching becomes correlated over long distances, and it is conceivable that this small value of $R_{\mathrm{T}} \approx 8$ corresponds to the small number of domains over which the receptors effectively switch in concert. More work is needed to understand whether receptor switching near a critical point can effectively be described by an MWC model, and whether the design rule unveiled here (Eq. 5.29), also generalizes to a receptor array near a critical point. Lastly, further study is necessary to understand whether information transmission in this system is maximized near a critical point [93].

The adaptation time of the *E. coli* chemotaxis system has repeatedly been shown to be ~ 10s, yielding $\tau_{\mathrm{m}}^{\exp}/\tau_v \approx 8$ [19, 20, 62]. Given the estimated resource allocation in the effective MWC description, $X_{\mathrm{T}} = 10^4$ and $R_{\mathrm{T}} = 8$ with $N_{\mathrm{r}} = 12$ (Section 4.5 and Appendix 4.C), this adaptation time is close to optimal for gradient steepnesses $g \lesssim 4\mathrm{mm}^{-1}$ (Fig. 5.4). In particular, while decreasing the methylation time improves the prediction accuracy in steeper gradients, it reduces information transmission in shallower gradients. On the other hand, while increasing the methylation time beyond the measured one decreases the accuracy in steeper gradients, the improvement in shallow gradients is only very minor because the system is already very close to the fundamental bound on the predictive information as set by the resource constraint and the gradient steepness (Fig. 5.4). These arguments show that the methylation time of *E. coli* is indeed optimal for sensing shallow gradients with $g \lesssim 4\mathrm{mm}^{-1}$. It suggests that the chemotaxis system has been optimized for navigating weak gradients. To get an idea of what this gradient steepness means we can compare it to the length of an *E. coli* cell, which is ~ 1μm. To cover a gradient length scale $g^{-1} = 1/4$mm thus requires the cell to move at least 250 times its body length, corresponding to approximately 10 runs in the same direction [94–96]. This illustrates how extremely shallow the gradients that *E. coli* can encounter likely are. Moreover, it suggests that it is most important to maximize accuracy in shallow gradients, where it is hard to distinguish signal from noise. In steeper gradients *E. coli* would be further from the optimal design, but the total information it obtains about the signal of interest is still larger because the input fluctuations are bigger. Clearly, a network that would adapt the adaptation time to the steepness of the gradient, would probably perform better over a broader range of gradient steepness. However, such a network would be (much) more complicated with a higher resource cost. It thus appears that evolution has optimized the network for sensing those gradients that are most difficult to detect.

5.8. Changing the signal statistics

Up to this point we have assumed that the cell navigates environments in which the nutrient concentration varies much more widely than the change in concentration over the duration of a run (Section 5.2). Indeed, we expect that this is the most natural regime for many microorganisms and especially for *E. coli*, as it is clear that *E. coli* can indeed measure concentration changes over a wide range of background concentrations based on the dissociation constants of the receptor in its active and inactive state, which differ by two orders of magnitude [56–58]. However, to illustrate the effect that the choice of signal statistics has on our results, we here consider a scenario in which the concentration and its change during a run are of the same order. We may view this as a scenario in which the cell remains close to a constant nutrient peak.

Specifically, we consider a signal with the dynamics given in Eqs. 5.12 and 5.13, but with a concentration standard deviation $\sigma_\ell = 2\tau_v\sigma_v$, yielding $\omega_0 = (2\tau_v)^{-1}$. For these parameters, both the current concentration and the current derivative can be informative of the future derivative v_τ, depending on the forecast interval (Fig. 5.5(a), Eqs. 5.73 and 5.74).

In Chapter 2 we have shown that in this scenario, where both the current concentration value and the current concentration derivative are informative about the future derivative, the optimal network is one that bases its prediction on both the concentration and the derivative, proportional to the magnitude of their correlations with the signal of interest (Eq. 2.76). Our model of the chemotaxis network does not allow this, as it is set-up to exhibit perfect adaptation, and thus only measures the concentration change. However, in the limit of infinitely slow adaptation $\tau_\mathrm{m} \to \infty$, the network effectively becomes a push-pull network and measures the concentration only. Therefore, it is still interesting to investigate the optimal adaptation time that minimizes the relative error (SNR^{-1}) under these markedly different signal statistics.

To compute the SNR^{-1}, we derive the dynamic gain and the dynamical error for this signal, with $\sigma_\ell = 2\tau_v\sigma_v$ (Eqs. 5.78 and 5.83). Since we only change the signal statistics, the sampling error in estimating the receptor activity ($\sigma^{2,\mathrm{samp}}_{\hat{p}_{\tau_\mathrm{r}}}$, Eq. 5.22) and optimal resource allocation (Eq. 5.29) remain unchanged. We can then compute the relative prediction error (SNR^{-1}) of the chemotaxis network under the new signal statistics in the same manner as before (via Eqs. 5.2 and 5.11).

Figure 5.5(b) reveals that the prediction error now exhibits two distinct regions as a function of the adaptation time τ_m and the forecast interval τ, separated by a boundary on which the error diverges (dashed black line). These two regions reflect two distinct prediction strategies. For short forecast intervals the future derivative v_τ is more correlated with the current derivative $v(t_0)$, than with the current value $\ell(t_0)$, i.e. $\rho_{vv}(\tau) > |\rho_{\ell v}(\tau)|$ [Fig. 5.5(a)]. In this regime, a short adaptation time allows the network to predict the future derivative v_τ based on the current, i.e. instantaneous, derivative $v(t_0)$. As the forecast interval increases, the contribution from the dynamical error rises more strongly than that from the sampling error (Eq. 5.84), which tends to decreases the optimal adaptation time. Simultaneously however, the future derivative also starts to become more strongly correlated to the current concentration value $\ell(t_0)$ as the forecast interval increases (Fig. 5.5a, blue curve). As a result, in the regime of large forecast intervals beyond the dashed black curve in Fig. 5.5(b), the optimal adaptation time diverges: the network effectively becomes a push-pull network and bases its prediction on the concentration value rather than its derivative.

Figure 5.5(b) shows that the forecast interval at which the error diverges becomes shorter for longer adaptation times (black dashed line). The prediction error diverges when the dynamic gain (Eq. 5.78), which quantifies the covariance between the current network output and the future signal derivative, becomes zero (Eq. 5.79). The point at which the dynamic gain becomes zero depends on the forecast interval τ via the signal correlation function $\rho_{vv}(\tau)$ (red curve Fig. 5.5(a), Eq. 5.73), which quantifies how much the current concentration derivative $v(t_0)$ is correlated with the future concentration derivative v_τ at the forecast interval τ. It also depends on the adaptation time τ_m, because that determines the degree to which the signal derivative taken by the network reflects the current derivative. When $\tau_\mathrm{m} \to 0$, the system takes an instantaneous derivative; the prediction error then diverges when the forecast interval $\tau = 2\tau_v$, precisely because the current derivative is then not correlated with the future derivative at that later time τ ($\rho_{vv}(2\tau_v) = 0$, see Fig. 5.5(a) red curve). For a longer adaptation time, the network takes a derivative over the signal further back into the past, i.e. a derivative centered around a time $\sim t_0 - \tau_\mathrm{m}$, which, as can be inferred from $\rho_{vv}(\tau)$ in Fig. 5.5(a) exploiting stationarity, is uncorrelated with the future derivative at a time $\sim t_0 - \tau_\mathrm{m} + 2\tau_v$.

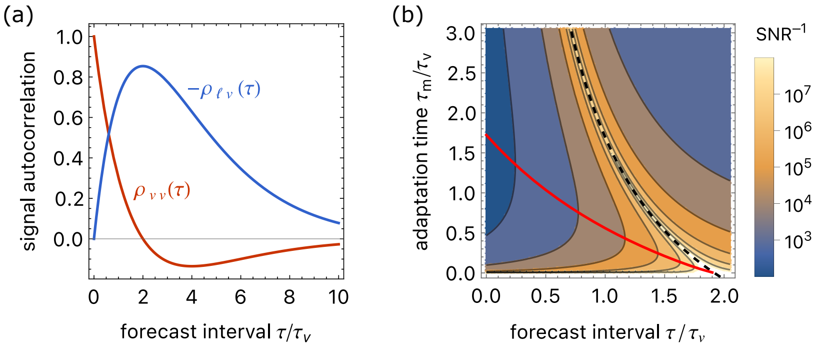


Figure 5.5: Two distinct regimes appear for the chemotaxis network predicting a future derivative when the signal concentration variations are of the same order as the change in concentration during a run: $\sigma_\ell = 2\tau_v\sigma_v$ as opposed to $\sigma_\ell \gg \tau_v\sigma_v$ (Fig. 5.3). (a) The correlation of the current signal derivative $v(t_0)$ (red curve, Eq. 5.73) and of the current signal concentration $\ell(t_0)$ (blue curve, Eq. 5.74) with the future derivative v_τ. (b) The SNR^{-1} (Eq. 5.84) is shown as a function of the adaptation time τ_m/τ_v and the forecast interval τ/τ_v for the chemotaxis network sensing a signal with statistics shown in (a). The optimal adaptation time in the regime of short forecast intervals and adaptation times (red curve) was computed by numerically minimizing Eq. 5.84. The error diverges where the dynamical gain is zero (Eq. 5.79), because the correlation between the network output and the future derivative vanishes (dashed black curve). Parameters other than σ_ℓ are as in Fig. 5.3 and code to reproduce the figure is available at [91].

5.9. DISCUSSION

Microorganisms that navigate chemical gradients need to predict the change in the concentration that they will encounter. For simple input signals where the change in concentration is Markovian, the optimal way to achieve this is to measure the current time derivative of the concentration [38]. Measuring such temporal concentration changes requires perfect adaptation. Moreover, to measure the most recent concentration change, the adaptation time must be short relative to the correlation time of the input. However, building and maintaining a biochemical network costs physical resources. When the resource availability is limited, the signal is obscured by noise in the network. The only way to lift the signal above the noise in this regime, is to increase the adaptation time. This trade-off between lifting the signal above the noise, and measuring a concentration change which is informative of the future input, sets the optimal adaptation time.

The optimal adaptation time depends on the amount of resources available to maintain the network, and the magnitude of changes in the input. The latter is set by the swimming behavior of the cell and the steepness of the chemical gradient it navigates. In steeper gradients the input changes more strongly, which reduces the sampling error and increases the signal-to-noise ratio. A smaller sampling error allows for a shorter adaptation time, which mitigates the dynamical error and maximizes the overall accuracy. Therefore, the optimal adaptation time to predict the concentration change scales approximately inversely with the gradient steepness. Interestingly, simulations show that the optimal adaptation time that maximizes navigational performance also increases as the gradient becomes more shallow [97, 98]. This indicates that predicting the concentration change is indeed important for successful navigation, in line with results of agent-based simulations on the interplay between prediction and navigation [7].

Our results provide a possible explanation for a puzzling observation. During chemotaxis, *E. coli* performs subsequent runs of approximately one second in different directions. Runs in the correct direction relative to the gradient are extended, and vice versa, such that the cell moves up a gradient of attractant on average. To implement this strategy, *E. coli* must predict how the concentration will change while it navigates the gradient. To this end, it seems natural to measure the change in concentration over the course of one run, i.e. over approximately one second. However, the adaptation time of *E. coli* is around ten seconds [19, 20, 62]. This raises the question, why would *E. coli* measure concentration changes over a timescale that is much longer than that of a run? Our work shows that the adaptation time must be this long to discern the signal from the inevitable biochemical signaling noise in shallow gradients.

Our analysis highlights that the optimal design of any sensing system depends on the signal statistics. In Chapter 4 we argued that a) *E. coli* aims to predict the future derivative; b) it can measure concentration changes over a range of background concentrations that is much larger than the typical concentration change during a run; c) in this regime, the optimal signaling system for predicting the derivative is a perfectly adaptive system, as *E. coli* has; d) its long adaptation time of $\tau_{\mathrm{m}} \approx 10$s is optimal for sensing shallow gradients; in fact, in this regime the predictive power of the *E. coli* chemotaxis system becomes exceedingly close to that of the optimal system. These observations together strongly suggest that *E. coli* has been optimized to sense shallow gradients, which is perhaps not so surprising since these gradients are the hardest to detect. Our previous analysis also showed that, since in this regime of shallow gradients the dynamics of the concentration derivative is Markovian, the optimal design does not depend on the forecast interval (Appendix 4.A). Our current analysis of a different input signal (Fig. 5.5) shows however, that in general, the optimal design depends on the signal statistics and the forecast interval. What the relevant signal statistics and forecast interval are depends on the dynamics of the environment and on how the organism navigates the environment [99, 100]. We leave these questions for future work.

More generally, our results provide insight into the optimal design of adaptive signaling networks. First and foremost, this improves our understanding of navigation behavior of microorganisms. But the uncovered principles might well hold more generally and shed light on other adaptive signaling networks as well, e.g. that of rod cells in the vertebrate eye [101]. Moreover, our theory facilitates the optimal design of microrobots that need to navigate environments without a map.

Appendix

5.A. The prediction error

Here we derive the general expression for the prediction error $\sigma^2_{\hat{p}_{\tau_\mathrm{r}}}$, which shows how the complete error decomposes into independent parts caused by fluctuations in the number of samples N, the error of a sampling process with a fixed number of samples and a constant input, and uninformative fluctuations from the input signal. Our starting point is the decomposition of the error in Eq. 5.7.

The first term of Eq. 5.7 is straightforward to compute, using the definition of $\hat{p}_{\tau_r}$ from Eq. 5.5 we obtain,

$$\mathrm{Var}\left(\mathbb{E}\left[\hat{p}_{\tau_r}|N\right]\right) = \mathrm{Var}\left(\mathbb{E}\left[\frac{1}{\overline{N}}\sum_{i=1}^{N} n_i(t_i)|N\right]_{t_i,n_i}\right)_N, \tag{5.32}$$

$$= \frac{1}{\overline{N}^2}\mathrm{Var}\left(N\mathbb{E}\left[\langle n(t_i)\rangle\right]_{t_i}\right)_N, \tag{5.33}$$

$$= \frac{p^2}{\overline{N}}, \tag{5.34}$$

where the subscripts after the expected values and variances denote the random variables over which the expectation is taken. For instance, in Eq. 5.32 the expected value is taken under a fixed number of samples N over the state $n_i \in \{0,1\}$ of each cluster, later also denoted with angle brackets as an ensemble average, and over all sampling times t_i, which are exponentially distributed with PDF [13]

$$f(t_i) = \frac{1}{\tau_\mathrm{r}} e^{-(t_0 - t_i)/\tau_\mathrm{r}}. \tag{5.35}$$

From Eq. 5.33 to Eq. 5.34 we use that the average number of active receptor clusters is defined as $p \equiv \mathbb{E}\left[\langle n(t_i)\rangle\right]_{t_i}$, which is constant with respect to N. The variance is subsequently taken over the Poisson distributed number of samples N, with both mean and variance $\overline{N}$. The resulting expression (Eq. 5.34) is the error that arises because the network cannot distinguish between those readout molecules that sampled an inactive cluster, and those that did not sample a cluster at all [13, 29].

We decompose the second term of Eq. 5.7 in two steps. First, we use the definition of $\hat{p}_{\tau_r}$ (Eq. 5.5) and split the self- and cross-terms in the covariance of the kinase activity:

$$\mathbb{E}\left[\mathrm{Var}\left(\hat{p}_{\tau_r}|N\right)\right] = \mathbb{E}\left[\mathrm{Var}\left(\frac{1}{\overline{N}}\sum_{i=1}^{N} n_i(t_i)|N\right)_{n_i,t_i}\right]_N, \tag{5.36}$$

$$= \frac{1}{\overline{N}^2}\mathbb{E}[N\mathrm{Var}\left(n_i(t_i)\right) + N(N-1)\mathrm{Cov}\left(n_i(t_i), n_j(t_j)\right)]_N, \tag{5.37}$$

$$= \frac{p(1-p)}{\overline{N}} + \mathrm{Cov}\left(n_i(t_i), n_j(t_j)\right). \tag{5.38}$$

From Eq. 5.37 to Eq. 5.38 we used that both the variance of each cluster and the covariance between clusters are independent of the number of samples N, and that for a

Poisson distributed number of samples N we have $\mathbb{E}[N(N-1)] = \overline{N}^2$. To continue, the covariance between different kinases at different times can be decomposed into contributions from the receptor noise, and fluctuations in the full history of the input signal, the trajectory $\boldsymbol{s}$,

$$\begin{aligned}\mathrm{Cov}\left(n_i(t_i)n_j(t_j)\right) &= \mathbb{E}\left[\mathrm{Cov}\left(n_i(t_i), n_j(t_j)|\boldsymbol{s}\right)\right]_{t_i,t_j,\boldsymbol{s}} + \mathrm{Cov}\left(\mathbb{E}[\langle n_i(t_i)|\boldsymbol{s}\rangle]_{t_i}, \mathbb{E}\left[\langle n_j(t_j)|\boldsymbol{s}\rangle\right]_{t_j}\right)_{\boldsymbol{s}}, \quad (5.39)\\ &= \mathbb{E}\left[\mathrm{Cov}\left(n_i(t_i), n_j(t_j)|\boldsymbol{s}\right)\right]_{t_i,t_j,\boldsymbol{s}} + \mathrm{Var}\left(\mathbb{E}[\langle n(t_i)|\boldsymbol{s}\rangle]_{t_i}\right)_{\boldsymbol{s}}, \quad (5.40)\end{aligned}$$

where we use that $\mathbb{E}[\langle n_i(t_i)|\boldsymbol{s}\rangle]_{t_i} = \mathbb{E}\left[\langle n_j(t_j)|\boldsymbol{s}\rangle\right]_{t_j}$. The two terms on the RHS of Eq. 5.40 respectively describe the covariance between clusters when the input is fixed, and the variance that is caused by input fluctuations. The first term is the receptor-level noise, which for the chemotaxis model considered in this work arises only from methylation Eqs. 5.58-5.63. The second term is the variance of the mean activity conditional on the input, which is the signal-induced variance. This signal induced variance comprises all variance caused by the input, so both the dynamical error and the variance that is informative of the signal of interest $\tilde{g}^2\sigma^2_{s_\tau}$ Eqs. 5.64-5.69.

Combining Eqs. 5.38 and 5.40 gives

$$\mathbb{E}\left[\mathrm{Var}\left(\hat{p}_{\tau_r}|N\right)\right] = \frac{p(1-p)}{\overline{N}} + \mathbb{E}\left[\mathrm{Cov}\left(n_i(t_i), n_j(t_j)|\boldsymbol{s}\right)\right]_{t_i,t_j,\boldsymbol{s}} + \mathrm{Var}\left(\mathbb{E}[\langle n(t_i)|\boldsymbol{s}\rangle]_{t_i}\right)_{\boldsymbol{s}}. \quad (5.41)$$

Finally, the third term of Eq. 5.7 is the contribution of the signal of interest to the output variance:

$$\begin{aligned}\mathrm{Var}\left(\mathbb{E}\left[\hat{p}_{\tau_r}|s_\tau\right]\right) &= \mathrm{Var}\left(\mathbb{E}\left[\frac{1}{\overline{N}}\sum_{i=1}^{N} n_i(t_i)|s_\tau\right]_{t_i,n_i,N}\right)_{s_\tau}, \quad (5.42)\\ &= \mathrm{Var}\left(\mathbb{E}[\langle n(t_i)|s_\tau\rangle]_{t_i}\right)_{s_\tau}, \quad (5.43)\\ &= \tilde{g}^2\sigma^2_{s_\tau}, \quad (5.44)\end{aligned}$$

where in the last step we have used the dynamic input output relation of Eq. 5.1. The dynamic gain $\tilde{g}$ of the chemotaxis network is derived in Eqs. 5.54-5.57. Substituting the equalities of Eqs. 5.34, 5.41 and 5.44 in Eq. 5.7 of the main text gives the complete prediction error given in Eq. 5.11 in the main text.

We note that this derivation deviates from that of Malaguti and Ten Wolde [29] in that Eq. 5.36 includes the contributions from all signal variations, including the informative signal variations (which are then subtracted from the full variance in Eq. 5.7), while in [29] the corresponding term does not contain these informative signal fluctuations. While the final result is identical, the derivation presented here is arguably easier.

5.B. THE CHEMOTAXIS NETWORK

In the *E. coli* chemotaxis network, receptors cooperatively control the activity of the kinase CheA, and the activity is adaptive due to the methylation of inactive receptors

[19, 60–64]. We here follow the widely used approach to describe the effects of receptor cooperativity and methylation on kinase activity via the Monod-Wyman-Changeux (MWC) model [20, 50, 52, 57, 58, 62, 65–67]. In this model, each receptor can switch between an active and inactive conformational state n and receptors are partitioned into clusters of equal size N_r. In the spirit of the MWC model, receptors within a cluster switch conformation in concert, so that each cluster is either active or inactive [65]. Furthermore, it is assumed that receptor-ligand binding and conformational switching are faster than the other timescales in the system, such that the activity state of the receptor can effectively be described by its equilibrium probability to be active, given the methylation level of the cluster m and the external ligand concentration ℓ. The probability for the receptor cluster to be active is then described by:

$$a(\ell, m) \equiv \langle n|\ell, m\rangle = (1+\exp(\Delta F_T(\ell, m)))^{-1}, \tag{5.45}$$

where $\Delta F_T(\ell, m) = -\Delta E_0 + N_\mathrm{r}(\Delta F_\ell(\ell) + \Delta F_m(m))$ is the free-energy difference between the active and inactive state, which is a function of free-energy difference arising from ligand binding and methylation:

$$\Delta F_\ell(\ell) = \ln(1+\ell(t)/K_\mathrm{D}^\mathrm{I}) - \ln(1+\ell(t)/K_\mathrm{D}^\mathrm{A}), \tag{5.46}$$

$$\Delta F_m(m) = \tilde{\alpha}(\bar{m} - m(t)). \tag{5.47}$$

Between the two states the cluster has an altered dissociation constant, which is denoted K_D^I for the inactive state, and K_D^A for the active state. The free-energy difference due to methylation has been experimentally shown to depend approximately linearly on the methylation level [62]. We assume that inactive receptors are irreversibly methylated, and active receptors irreversibly demethylated, with zero-order ultrasensitive kinetics [36, 58, 68]. The methylation dynamics of a receptor cluster is then given by:

$$\dot{m} = (1-a(\ell, m))k_R - a(\ell, m)k_B + B_m(a)\xi(t), \tag{5.48}$$

with $B_m(a) = \sqrt{(1-a(\ell, m))k_R + a(\ell, m)k_B}$, and unit white noise $\xi(t)$. These dynamics give rise to perfect adaptation, since from this equation we find that the steady state cluster activity is given by $p \equiv \bar{a} = 1/(1+k_B/k_R)$, thus indeed independent of the ligand concentration.

In this work we consider linear dynamics, we therefore employ a linear noise approximation [30]. The deviation of the equilibrium cluster activity from its mean $\delta a(t) = a(t) - p$ is then given by

$$\delta a(t) \equiv \langle n(t)|\delta\ell, \delta m\rangle - p = \alpha\delta m(t) - \beta\delta\ell(t), \tag{5.49}$$

with $\alpha = \tilde{\alpha}N_\mathrm{r}p(1-p)$ and $\beta = \kappa N_\mathrm{r}p(1-p)$, with $\kappa = (1+K_\mathrm{D}^\mathrm{I}/c_0)^{-1} - (1+K_\mathrm{D}^\mathrm{A}/c_0)^{-1}$. For the methylation dynamics on one cluster we then obtain,

$$\dot{\delta m} = -\delta a(t)/(\alpha\tau_\mathrm{m}) + \eta_m(t), \tag{5.50}$$

where we have introduced the adaptation time $\tau_\mathrm{m} = (\alpha(k_R+k_B))^{-1}$ and $\eta_m(t)$ is Gaussian white noise on a single cluster with correlation function

$$\left\langle \eta_{m_i}(t)\eta_{m_j}(t')\right\rangle = \delta_{ij}\delta(t-t')\frac{2p(1-p)}{\alpha\tau_\mathrm{m}} \tag{5.51}$$

between the i^{th} and j^{th} receptor cluster, where δ_{ij} is the Kronecker delta. Combining Eqs. 5.49 and 5.50 yields the change in activity over time

$$\dot{\delta a} = -\delta a(t)/\tau_{\mathrm{m}} - \beta v(t) + \alpha \eta_m(t), \tag{5.52}$$

where we have the change in concentration over time $v(t) \equiv \dot{\delta\ell}$. Using Eq. 5.52 we can also express the instantaneous activity as

$$\delta a(t) = \int_{-\infty}^{t} dt' \left(\alpha\eta_m(t') - \beta v(t')\right) e^{-(t-t')/\tau_{\mathrm{m}}}. \tag{5.53}$$

This expression shows that the cluster activity, when we average out the methylation noise, reflects the change in concentration weighted exponentially over the past adaptation time τ_{m}.

DYNAMIC GAIN

The dynamic gain of the network can be obtained by deriving the average response of the network to the signal of interest s_τ. In general we have the expression given in Eq. 5.1 for the dynamic input output relation of linear signaling networks. In our case the signal of interest is the future concentration derivative $s_\tau = v_\tau$. Using Eqs. 5.49 and 5.53, we find for the average conditional activity,

$$\langle n(t_i)|v_\tau\rangle = \mathbb{E}[\langle n(t_i)|v_\tau,\delta\ell,\delta m\rangle]_{\delta\ell,\delta m}, \tag{5.54}$$

$$= p - \beta\int_{-\infty}^{t_i} dt \langle v(t)|v_\tau\rangle e^{-(t_i-t)/\tau_{\mathrm{m}}}, \tag{5.55}$$

$$= p - e^{-(t_0+\tau-t_i)/\tau_v} \frac{\tau_{\mathrm{m}}\beta v_\tau}{1+\tau_{\mathrm{m}}/\tau_v} \tag{5.56}$$

where we used that the conditional mean derivative is $\langle v(t)|v_\tau\rangle = v_\tau \exp(-(t_0+\tau-t)/\tau_v)$, also see Eq. 5.14. Averaging over all sampling times, distributed as in Eq. 5.35, gives

$$\mathbb{E}[\langle n(t_i)|v_\tau\rangle]_{t_i} = p - \frac{\tau_{\mathrm{m}}\beta e^{-\tau/\tau_v} v_\tau}{(1+\tau_{\mathrm{m}}/\tau_v)(1+\tau_{\mathrm{r}}/\tau_v)}. \tag{5.57}$$

Comparison to Eq. 5.1 yields the dynamic gain $\tilde{g}$ given in Eq. 5.18.

RECEPTOR NOISE

The variance that is caused by receptor-level (here methylation) noise is the covariance between clusters under a fixed input trajectory, i.e. the first term of Eq. 5.40. We can write this covariance in terms of the equilibrium activity as follows, using Eq. 5.49 and noting that $\delta\ell(t)$ is contained in $\boldsymbol{s}$ for $t \le t_0$:

$$\mathbb{E}\left[\mathrm{Cov}\big(n_i(t_i), n_j(t_j)|\boldsymbol{s}\big)\right]_{t_i,t_j,\boldsymbol{s}} = \mathbb{E}\left[\left\langle n_i(t_i)n_j(t_j)|\boldsymbol{s},\delta m\right\rangle\right]_{t_i,t_j,\boldsymbol{s},\delta m} - p^2, \tag{5.58}$$

$$= \mathbb{E}\left[\langle n_i(t_i)|\boldsymbol{s},\delta m\rangle \left\langle n_j(t_j)|\boldsymbol{s},\delta m\right\rangle - p^2\right]_{t_i,t_j,\boldsymbol{s},\delta m}, \tag{5.59}$$

$$= \mathbb{E}\left[\left\langle \delta a_i(t_i)\delta a_j(t_j)|\boldsymbol{s}\right\rangle\right]_{t_i,t_j,\boldsymbol{s},\delta m}. \tag{5.60}$$

In Eq. 5.58 we condition on- and average over δm to make the connection between the instantaneous cluster state n_i and the cluster activity a_i (Eq. 5.49). Then in Eq. 5.59 we use the fact that when conditioned on both the signal and the methylation level, the cluster states are independent. The covariance in the cluster activity conditioned on the full past input trajectory (Eq. 5.60) depends only on the methylation noise, using Eqs. 5.53 and 5.51 and keeping the sampling times fixed,

$$
\begin{aligned}
\mathbb{E}\left[\langle \delta a_i(t_i)\delta a_j(t_j)|\boldsymbol{s}\rangle\right]_{\boldsymbol{s},\delta m} &= \mathbb{E}\Big[\alpha^2 \int_{-\infty}^{t_i} dt \int_{-\infty}^{t_j} dt' \\
&\qquad \left\langle \eta_{m_i}(t)\eta_{m_j}(t')\right\rangle e^{-(t_i-t)/\tau_{\mathrm{m}}} e^{-(t_j-t')/\tau_{\mathrm{m}}}\Big]_{\boldsymbol{s},\delta m} && (5.61)\\
&= \left\langle \delta_{ij}\right\rangle \frac{2\alpha p(1-p)}{\tau_{\mathrm{m}}} \int_{-\infty}^{t^-} dt e^{-(t^- - t)/\tau_{\mathrm{m}}} e^{-(t^+ - t)/\tau_{\mathrm{m}}}, && (5.62)\\
&= \frac{\alpha p(1-p)}{R_{\mathrm{T}}} e^{-|t_i-t_j|/\tau_{\mathrm{m}}}, && (5.63)
\end{aligned}
$$

where $t^+ \equiv \max(t_i, t_j)$ and $t^- \equiv \min(t_i, t_j)$, and the number of receptor clusters R_{T} arises as the average Kronecker delta over all clusters: $\langle \delta_{ij}\rangle = 1/R_{\mathrm{T}}$. Averaging over the exponentially distributed sampling times t_i and t_j (both following Eq. 5.35), yields the receptor noise given in Eq. 5.21.

SIGNAL INDUCED CORRELATIONS

The covariance in the output caused by the variation in the past input signal is given by the second term of Eq. 5.40. It describes all variance in the output caused by input fluctuations, so it comprises both the dynamical error and the informative part $\tilde{g}^2\sigma^2_{s_\tau}$. We rewrite the instantaneous activity to the equilibrium activity using Eq. 5.49 and considering that $\delta\ell(t)$ is contained in $\boldsymbol{s}$ for $t \leq t_0$:

$$
\begin{aligned}
\mathrm{Var}\left(\mathbb{E}\left[\langle n(t_i)|\boldsymbol{s}\rangle\right]_{t_i}\right)_{\boldsymbol{s}} &= \mathrm{Var}\left(\mathbb{E}\left[\langle n(t_i)|\boldsymbol{s},\delta m\rangle\right]_{t_i,\delta m}\right)_{\boldsymbol{s}}, && (5.64)\\
&= \mathrm{Var}\left(p + \mathbb{E}\left[\langle \delta a(t_i)|\boldsymbol{s}\rangle\right]_{t_i,\delta m}\right)_{\boldsymbol{s}}, && (5.65)\\
&= \mathrm{Var}\left(-\frac{\beta}{\tau_{\mathrm{r}}}\int_{-\infty}^{t_0} dt_i \int_{-\infty}^{t_i} dt v(t) e^{-(t_i-t)/\tau_{\mathrm{m}}} e^{-(t_0-t_i)/\tau_{\mathrm{r}}}\right) && (5.66)
\end{aligned}
$$

where in Eq. 5.64 we again condition on- and average over δm to make the connection between $n(t)$ and $a(t)$ (Eq. 5.49). In Eq. 5.66 we used Eq. 5.53 and the sampling time distribution of Eq. 5.35. Using the correlation function of the concentration derivative, Eq. 5.14, we continue from Eq. 5.66 to obtain

$$
\begin{aligned}
&\mathrm{Var}\left(\mathbb{E}\left[\langle n(t_i)|\boldsymbol{s}\rangle\right]_{t_i}\right)_{\boldsymbol{s}} = \frac{\sigma_v^2\beta^2}{\tau_{\mathrm{r}}^2}\int_{-\infty}^{t_0} dt_i \int_{-\infty}^{t_0} dt_j \Big(\\
&\int_{-\infty}^{t_i} dt \int_{-\infty}^{t_j} dt' e^{-|t-t'|/\tau_v} e^{-(t_i-t)/\tau_{\mathrm{m}}} e^{-(t_j-t')/\tau_{\mathrm{m}}}\Big) e^{-(t_0-t_i)/\tau_{\mathrm{r}}} e^{-(t_0-t_j)/\tau_{\mathrm{r}}}. \qquad (5.67)
\end{aligned}
$$

First we perform the integrals over t and t', which yields,

$$\mathrm{Var}\left(\mathbb{E}\left[\langle n(t_i)|\boldsymbol{s}\rangle\right]_{t_i}\right)_{\boldsymbol{s}} = \frac{\sigma_v^2\beta^2/\tau_{\mathrm{r}}^2}{1/\tau_v^2 - 1/\tau_{\mathrm{m}}^2}\int_{-\infty}^{t_0} dt_i \int_{-\infty}^{t_0} dt_j \Big(\frac{\tau_{\mathrm{m}}}{\tau_v} e^{-|t_i-t_j|/\tau_{\mathrm{m}}} - e^{-|t_i-t_j|/\tau_v}\Big) e^{-(t_0-t_i)/\tau_{\mathrm{r}}} e^{-(t_0-t_j)/\tau_{\mathrm{r}}}. \tag{5.68}$$

Finally, computing the integrals over the sampling times t_i and t_j gives,

$$\mathrm{Var}\left(\mathbb{E}\left[\langle n(t_i)|\boldsymbol{s}\rangle\right]_{t_i}\right)_{\boldsymbol{s}} = \frac{\tau_{\mathrm{m}}^2\beta^2\sigma_v^2(1+\tau_{\mathrm{r}}/\tau_{\mathrm{m}}+\tau_{\mathrm{r}}/\tau_v)}{(1+\tau_{\mathrm{m}}/\tau_v)(1+\tau_{\mathrm{r}}/\tau_v)(1+\tau_{\mathrm{r}}/\tau_{\mathrm{m}})}, \tag{5.69}$$

which is equivalent to the expression in main text Eq. 5.25 with the static gain of Eq. 5.19.

5.C. Optimal adaptation time

Here we give a comprehensive derivation of the approximate optimal adaptation time. To gain analytical insight into the optimal adaptation time we first consider that the adaptation time τ_{m} must be larger than the response time τ_{r} to yield a non-zero response to transient input changes. Subsequently taking the derivative of Eq. 5.27 with respect to τ_{m} then gives,

$$\frac{\partial \mathrm{SNR}^{-1}}{\partial \tau_{\mathrm{m}}} = \frac{e^{2\tau/\tau_v}}{\tau_v}\left(1+\frac{\tau_{\mathrm{r}}}{\tau_v}\right)^2\left[1-\frac{2\,(1+\tau_v/\tau_{\mathrm{m}})}{(\tau_{\mathrm{m}}\beta g\sigma_{v_x})^2}\left(\frac{p^2}{\overline{N}}+\frac{p(1-p)}{\overline{N}_{\mathrm{I}}}\right)\right], \text{ for } \tau_{\mathrm{m}} \gg \tau_{\mathrm{r}}, \tag{5.70}$$

where the number of independent samples $\overline{N}_{\mathrm{I}}$ is given by Eq. 5.23 with $R_{\mathrm{I}} = R_{\mathrm{T}}/\alpha$. Now considering that for *E. coli* the adaptation time is much larger than the signal correlation time gives, up to the prefactor,

$$\frac{\partial \mathrm{SNR}^{-1}}{\partial \tau_{\mathrm{m}}} \propto 1-\frac{2}{(\tau_{\mathrm{m}}\beta g\sigma_{v_x})^2}\left(\frac{p^2}{\overline{N}}+\frac{p(1-p)}{\overline{N}_{\mathrm{I}}}\right), \tag{5.71}$$

for $\tau_{\mathrm{m}} \gg \tau_{\mathrm{r}}, \tau_v$. Equating Eq. 5.71 to zero and solving for $\tau_{\mathrm{m}}^{\mathrm{opt}}$ yields one positive solution, given in Eq. 5.31.

Inspection of Eq. 5.31 also reveals that the limiting resource of the network will set the optimal adaptation time. This becomes apparent when we express $\overline{N}$ (Eq. 5.6) and $\overline{N}_{\mathrm{I}}$ (Eq. 5.23) in terms of the number of readout molecules X_{T} and receptor clusters R_{T}. Substitution in Eq. 5.31 then yields for the analytical approximation of the optimal adaptation time

$$\tau_{\mathrm{m}}^{\mathrm{opt}} \approx \frac{\sqrt{2}}{\beta g\sigma_{v_x}}\sqrt{\frac{p^2}{X_{\mathrm{T}}f(1-f)}+\frac{\alpha p(1-p)}{R_{\mathrm{T}}}}. \tag{5.72}$$

When $R_{\mathrm{T}} \gg X_{\mathrm{T}}$ the receptor noise (final RHS term in Eq. 5.72, also see Eq. 5.21) becomes negligibly small, and the adaptation time is set by the number of readout molecules X_{T}, which set the number of receptor samples $\overline{N}$ (Eq. 5.6) and thus the first two terms of the sampling error in Eq. 5.11. Figure 5.6(a) illustrates how the optimal adaptation time

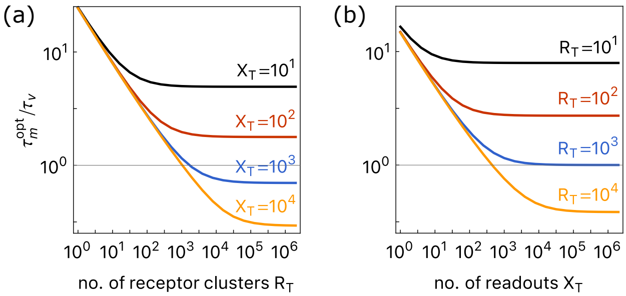


Figure 5.6: The limiting resource sets the optimal adaptation time. (a) The optimal adaptation time, determined by numerically minimizing Eq. 5.27, as a function of the number of receptor clusters R_T, for different numbers of readout molecules X_T. (b) The optimal adaptation time as a function of the number of readout molecules X_T, for different numbers of receptor clusters R_T. The panels show that X_T and R_T limit the prediction error like weak links in a chain, as observed for a simple push-pull network without methylation feedback [13, 15, 29]. Parameters are as in Fig. 5.3 and code to reproduce the figure is available at [91].

becomes independent of R_T when X_T is limiting. Vice versa, when $X_T \gg R_T$, the optimal adaptation time is set by the receptor noise [Fig. 5.6(b)]. The panels of Fig. 5.6 show that the receptors and the readout molecules limit sensing like weak links in a chain: the prediction error, and concomitantly the optimal adaptation time, is set by the limiting resource; the error cannot be lowered by increasing the other resource [13, 15, 29].

The observation that the limiting resource sets the optimal adaptation time can be understood when we consider that the optimal adaptation time arises from a trade-off between the sampling error and the dynamical error (Eq. 5.27). To minimize the dynamical error, the adaptation time τ_m must be zero, such that the network output reflects the most recent signal derivative. This is analogous to the dynamical error for the push-pull network studied by Malaguti and Ten Wolde [15], which is minimized by reducing the integration time τ_r to zero such that the network output reflects the most recent signal concentration. Importantly, unlike the sampling error, the dynamical error does not depend on R_T and X_T, but only on timescales. To reduce the sampling error of the chemotaxis network, the adaptation time must increase. Again, this seems similar to the role of the integration time in the push-pull network, where an increase in the integration time can also reduce the sampling error [15]. However, in the push-pull network, increasing the integration time can reduce the sampling error because it enables the network to time-average the noise arising from receptor-ligand binding. This also means that in the push-pull network, the optimal integration time reduces to zero when $R_T \gg X_T$, because time averaging means that more receptor samples or concentration measurements are taken *per* receptor, which requires $X_T > R_T$. This is markedly different from the chemotaxis network [Fig. 5.6(a)], because the role of the adaptation time is fundamentally different from that of the integration time. Indeed, increasing the adaptation time reduces the sampling error because it increases the dynamical gain (Eq. 5.18). This enables the network to lift the signal above the noise.

5.D. CHANGING SIGNAL STATISTICS

To illustrate the effect of the signal statistics on the prediction error we consider a signal with the dynamics given in Eqs. 5.12 and 5.13, but with a concentration standard deviation $\sigma_\ell = 2\tau_v\sigma_v$, yielding $\omega_0 = (2\tau_v)^{-1}$. These statistics describe a critically damped harmonic oscillator. For such a signal, both the current concentration value and its derivative can be correlated to the future derivative v_τ, depending on the forecast interval:

$$\langle \delta v(t_0)\delta v(\tau)\rangle = \sigma_v^2\left(1-\frac{\tau}{2\tau_v}\right)e^{-\tau/(2\tau_v)}, \tag{5.73}$$

$$\langle \delta \ell(t_0)\delta v(\tau)\rangle = -\sigma_v^2\tau e^{-\tau/(2\tau_v)}. \tag{5.74}$$

These correlation functions are shown in Fig. 5.5(a).

The dynamic gain quantifies the covariance between the network output and the signal of interest, and therefore naturally depends on the signal statistics. To derive the dynamic gain under the new signal statistics we follow the same procedure as before, starting from Eq. 5.54 and using that now $\langle v(t)|v_\tau\rangle = v_\tau[1-\tau/(2\tau_v)]\exp[-\tau/(2\tau_v)]$ (see Eq. 5.73), we obtain

$$\langle n(t_i)|v_\tau\rangle = p - \beta\int_{-\infty}^{t_i} dt\,\langle v(t)|v_\tau\rangle\, e^{-(t_i-t)/\tau_\mathrm{m}}, \tag{5.75}$$

$$= p - e^{-(t_0+\tau-t_i)/(2\tau_v)}\frac{\tau_\mathrm{m}\beta v_\tau}{1+\tau_\mathrm{m}/(2\tau_v)}\left(1-(t_0+\tau-t_i)/(2\tau_v)-\frac{1}{1+2\tau_v/\tau_\mathrm{m}}\right). \tag{5.76}$$

Averaging over all sampling times t_i (Eq. 5.35) yields

$$\mathbb{E}[\langle n(t_i)|v_\tau\rangle]_{t_i} = p + \tilde{g}v_\tau, \tag{5.77}$$

with the dynamic gain

$$\tilde{g} = \frac{-\tau_\mathrm{m}\beta e^{-\tau/(2\tau_v)}}{(1+\tau_\mathrm{m}/(2\tau_v)(1+\tau_\mathrm{r}/(2\tau_v))}\left(1-\frac{\tau}{2\tau_v}-\frac{1}{1+2\tau_v/\tau_\mathrm{m}}-\frac{1}{1+2\tau_v/\tau_\mathrm{r}}\right). \tag{5.78}$$

From this expression we can see that the dynamic gain goes to zero when

$$\frac{\tau}{\tau_v} = 2\left(1-\frac{1}{1+2\tau_v/\tau_\mathrm{m}}-\frac{1}{1+2\tau_v/\tau_\mathrm{r}}\right). \tag{5.79}$$

When this relation holds, the current network output and the future signal derivative are not correlated. When $\tau_\mathrm{m}\to 0$ the network takes an instantaneous derivative; indeed, in this limit the dynamical gain is zero when $\tau = 2\tau_v$ (assuming $\tau_\mathrm{r} \ll \tau_v$), because the current derivative is not correlated with the future derivative v_τ for this forecast interval τ, under these signal statistics (Eq. 5.73). The dependence on the adaptation time arises because the adaptation time determines to what extent the current network output reflects the most recent derivative, or an exponentially weighted average of derivatives further in the past (Eq. 5.53). Similarly, τ_r sets the timescale over which the receptor activity is averaged on the level of the readout, potentially introducing an additional delay. However, for the chemotaxis network typically $\tau_\mathrm{r} \ll \tau_v$, such that the contribution of the final RHS term in Eq. 5.79 is relatively minor.

To determine the dynamical error under the new signal statistics (Eqs. 5.73 and 5.74), we first derive the total variance in the output caused by input fluctuations (second term of Eq. 5.40). We follow the same procedure as in Eqs. 5.64-5.69, starting from Eq. 5.66 and using the derivative autocorrelation of Eq. 5.73 yields

$$
\begin{aligned}
\mathrm{Var}\left(\mathbb{E}\left[\langle n(t_i)|\boldsymbol{s}\rangle\right]_{t_i}\right)_{\boldsymbol{s}} &= \mathrm{Var}\left(-\frac{\beta}{\tau_\mathrm{r}}\int_{-\infty}^{t_0} dt_i \int_{-\infty}^{t_i} dt\, v(t) e^{-(t_i-t)/\tau_\mathrm{m}} e^{-(t_0-t_i)/\tau_\mathrm{r}}\right),\\
&= \frac{\sigma_v^2\beta^2}{\tau_\mathrm{r}^2}\int_{-\infty}^{t_0} dt_i \int_{-\infty}^{t_0} dt_j \Big(\int_{-\infty}^{t_i} dt \int_{-\infty}^{t_j} dt' e^{-|t-t'|/(2\tau_v)}\Big[\\
&\quad 1-\frac{|t-t'|}{2\tau_v}\Big] e^{-(t_i-t)/\tau_\mathrm{m}} e^{-(t_j-t')/\tau_\mathrm{m}}\Big) e^{-(t_0-t_i)/\tau_\mathrm{r}} e^{-(t_0-t_j)/\tau_\mathrm{r}}. \qquad (5.80)
\end{aligned}
$$

First we perform the integrals over t and t', which yields,

$$
\begin{aligned}
\mathrm{Var}\left(\mathbb{E}\left[\langle n(t_i)|\boldsymbol{s}\rangle\right]_{t_i}\right)_{\boldsymbol{s}} &= \frac{\sigma_v^2\beta^2/\tau_\mathrm{r}^2}{1/\tau_\mathrm{m}^2-1/(2\tau_v)^2}\int_{-\infty}^{t_0} dt_i \int_{-\infty}^{t_0} dt_j \Big(\frac{\tau_\mathrm{m}/\tau_v}{\tau_\mathrm{m}^2/(2\tau_v)^2-1} e^{-|t_i-t_j|/\tau_\mathrm{m}}\\
&+ e^{-|t_i-t_j|/(2\tau_v)}\Big[1-\frac{|t_i-t_j|}{2\tau_v}-\frac{2}{1-(2\tau_v)^2/\tau_\mathrm{m}^2}\Big]\Big) e^{-(t_0-t_i)/\tau_\mathrm{r}} e^{-(t_0-t_j)/\tau_\mathrm{r}}. \qquad (5.81)
\end{aligned}
$$

Finally, computing the integrals over the sampling times t_i and t_j gives,

$$
\mathrm{Var}\left(\mathbb{E}\left[\langle n(t_i)|\boldsymbol{s}\rangle\right]_{t_i}\right)_{\boldsymbol{s}} = \frac{\tau_\mathrm{m}^2\beta^2\sigma_v^2}{(1+\tau_\mathrm{m}/(2\tau_v))^2(1+\tau_\mathrm{r}/(2\tau_v))^2}\left(1+\frac{\tau_\mathrm{m}\tau_\mathrm{r}}{\tau_v(\tau_\mathrm{m}+\tau_\mathrm{r})}\right), \qquad (5.82)
$$

which describes all variance in the output caused by variations in the past input signal, i.e. it comprises both the dynamical error and the informative part $\tilde{g}^2\sigma_v^2$. Equation 5.82 happens to be very similar to the variance caused by a signal that is Markovian in its derivative (Eq. 5.69), the differences are that now the denominator of the prefactor is squared and the timescale τ_v has become $2\tau_v$ in the prefactor. Using Eq. 5.82 in Eq. 5.11 we obtain for the dynamical error

$$
\sigma_{\hat{p}_{\tau_\mathrm{r}}}^{2,\mathrm{dyn}} = \frac{\sigma_v^2\tau_\mathrm{m}^2\beta^2}{(1+\tau_\mathrm{m}/(2\tau_v))^2\,(1+\tau_\mathrm{r}/(2\tau_v))^2}\left(1+\frac{\tau_\mathrm{m}\tau_\mathrm{r}}{\tau_v(\tau_\mathrm{m}+\tau_\mathrm{r})}\right)-\tilde{g}^2\sigma_v^2. \qquad (5.83)
$$

This dynamical error for a signal with $\sigma_\ell = 2\tau_v\sigma_v$ increases monotonically in τ_m but, in contrast to the dynamical error for a signal with $\sigma_\ell \gg \tau_v\sigma_v$ (Eq. 5.26), it saturates as $\tau_\mathrm{m} \gg \tau_v$.

The absolute prediction error is $\sigma_{\hat{p}_{\tau_\mathrm{r}}}^2 = \sigma_{\hat{p}_{\tau_\mathrm{r}}}^{2,\mathrm{samp}} + \sigma_{\hat{p}_{\tau_\mathrm{r}}}^{2,\mathrm{dyn}}$, with the unchanged sampling error (Eq. 5.22) and the dynamical error of Eq. 5.83. Dividing the absolute prediction error by the dynamic gain (Eq. 5.78) and the signal variance σ_v^2 yields the relative error, or SNR^{-1} (also see Eq. 5.2):

$$
\begin{aligned}
&\mathrm{SNR}^{-1} =\\
&\frac{e^{\tau/\tau_v}}{\tau_\mathrm{m}^2\beta^2\sigma_v^2}\left(1+\frac{\tau_\mathrm{m}}{2\tau_v}\right)^2\left(1+\frac{\tau_\mathrm{r}}{2\tau_v}\right)^2\left(\frac{p^2}{\overline{N}}+\frac{p(1-p)}{\overline{N}_\mathrm{I}}\right)\Big/\left(1-\frac{\tau}{2\tau_v}-\frac{1}{1+2\tau_v/\tau_\mathrm{m}}-\frac{1}{1+2\tau_v/\tau_\mathrm{r}}\right)^2\\
&\quad + e^{\tau/\tau_v}\left(1+\frac{\tau_\mathrm{m}\tau_\mathrm{r}}{\tau_v(\tau_\mathrm{m}+\tau_\mathrm{r})}\right)\Big/\left(1-\frac{\tau}{2\tau_v}-\frac{1}{1+2\tau_v/\tau_\mathrm{m}}-\frac{1}{1+2\tau_v/\tau_\mathrm{r}}\right)^2-1. \qquad (5.84)
\end{aligned}
$$

The relative error is shown as a function of the forecast interval τ and the adaptation time τ_{m} in Fig. 5.5(b). When the dynamic gain vanishes (Eq. 5.79) the relative error diverges, which reflects that the network output then no longer contains information about the signal of interest.

6

ACCURACY OF THE GAUSSIAN INFORMATION RATE

MANUEL REINHARDT and AGE TJALMA

Efficient information processing is crucial for both living organisms and engineered systems. The mutual information rate, a core concept of information theory, quantifies the amount of information shared between the trajectories of input and output signals, and allows to quantify information flow in dynamic systems. A common approach for estimating the mutual information rate is the Gaussian approximation, which assumes that the input and output trajectories follow Gaussian statistics. However, this method is limited to linear systems, and its accuracy in nonlinear or discrete systems remains unclear. In this work, we assess the accuracy of the Gaussian approximation for non-Gaussian systems by leveraging Path Weight Sampling (PWS), a recent technique for exactly computing the mutual information rate. In two case studies, we examine the limitations of the Gaussian approximation. First, we focus on discrete linear systems and demonstrate that, even when the system's statistics are nearly Gaussian, the Gaussian approximation fails to accurately estimate the mutual information rate. Second, we explore a continuous diffusive system with a nonlinear transfer function, revealing significant deviations between the Gaussian approximation and the exact mutual information rate as nonlinearity increases. Our results provide a quantitative evaluation of the Gaussian approximation's performance across different stochastic models and highlight when more computationally intensive methods, such as PWS, are necessary.

The work presented in this chapter was performed by Manuel Reinhardt (PWS simulations) and Age Tjalma (LNA calculations) as shared first authors; manuscript for publication is in preparation.

For the functioning of both living and engineered systems it is paramount that they collect and process information effectively. Increasingly, it has become clear that beyond instantaneous properties, the dynamic features of an input signal or system output often encode valuable information [20, 35–37, 102, 103]. Prime examples in biology include bacterial chemotaxis, which responds to temporal changes in concentration [19], the transcription factor NF-κB, which encodes information about input signals in its dynamic response [104], and neuronal information processing, where information is encoded in the sequence and timing of spikes [105]. Beyond biology, dynamic input signals are critical for various sensing systems, such as those used in automated factories or self-driving cars.

To understand and evaluate the performance, potential improvements, and limitations of these systems in processing information, we need appropriate metrics that capture their full information processing capability. Information theory, introduced by Shannon [41], provides the most general mathematical framework for such metrics. The mutual information and mutual information rate measure how much one random variable reduces uncertainty about another, quantified in bits. It is relatively straightforward to quantify the information shared between scalar properties of the input and output, as has been done in various forms [10–12, 32, 38, 73, 83, 86, 106]. However, capturing all information in dynamical properties of the input and the output is much more challenging. To do so, one must consider the information encoded in time-varying trajectories of the variables of interest. Yet, due to the high dimensionality of the trajectory space, computing the mutual information between such trajectories is notoriously difficult.

A major advancement in this area has been the Gaussian approximation of the mutual information rate [36, 37], based on the assumption of input and output trajectories following jointly Gaussian statistics. This assumption makes it possible to compute the mutual information rate directly from the two-point correlation functions of the input and output. It is thus straightforward to apply the Gaussian approximation to experimental data. Moreover, given a mechanistic model of the underlying dynamics, the Gaussian approximation can be used to derive analytical expressions for the information rate [20, 36, 37]. Crucially however, the assumption of Gaussian statistics restricts the method to linear systems, as Gaussian statistics can only arise in such systems [40].

Understanding when the Gaussian approximation is accurate is critical because many real-world systems, such as biological and engineered sensory systems, exhibit nonlinear dynamics. Real-world systems often exhibit complex statistical behaviors such as bimodality, discrete jumps, or heavy tails, all of which deviate from purely Gaussian dynamics. Such non-Gaussian behavior typically results from intrinsic nonlinearities in the system, but determining the degree of a system's deviation from linearity is difficult [33, 107, 108], and the extent to which the approximation loses accuracy in nonlinear systems is unclear. Thus, although the Gaussian approximation offers a computationally simple method to estimate information transmission, it remains an open question under what conditions this approximation is sufficiently accurate.

Until recently, addressing this question has been hard because there was no reliable benchmark for the exact information rate. Without a method to compute the true information rate of a non-Gaussian system, it is impossible to rigorously assess the accuracy of the Gaussian approximation. This gap was filled recently by the development of two independent methods [35, 102] for computing the information rate accurately even in systems that significantly deviate from Gaussian behavior. Here we leverage one of these methods: Path Weight Sampling (PWS) [35]. This is a Monte Carlo technique which is an exact method for calculating the mutual information rate in a wide range of stochastic models.

Using PWS, we can directly evaluate the accuracy of the Gaussian approximation in models that exhibit explicit non-Gaussian features, and study the approximation's robustness in typical applications.

In this chapter, we investigate the accuracy of the approximate Gaussian information rate through two case studies. The first focuses on discrete linear systems, where the statistics are thus non-Gaussian due to the discrete nature of the network. Perhaps surprisingly, the Gaussian approximation fails to accurately estimate the mutual information rate in this case, even when the statistics are nearly Gaussian [35, 102]. We show that a recently developed reaction-based "discrete approximation" by Moor and Zechner [102] is much more accurate. This suggests that the Gaussian approximation fails because it cannot distinguish the individual reaction events.

The second case study examines a continuous diffusive system with a nonlinear transfer function. We demonstrate how intrinsic nonlinearity can cause significant deviations between the Gaussian approximation and the true mutual information rate. By varying both the degree of nonlinearity and the system's response timescale, we provide a comprehensive quantitative understanding of the Gaussian approximation's limitations in nonlinear systems. Additionally, we show that for such systems, the Gaussian approximation differs significantly when derived from empirical correlation functions compared to when it is analytically obtained from the nonlinear model, highlighting that the correct application of the approximation is important.

Our work translates into concrete recommendations on when to use which method for the computation of the information rate. It therefore enables researchers to more confidently determine when a simpler approximate method is sufficient, or when a more sophisticated method like PWS or the method developed by Moor and Zechner [102] should be used.

In what follows, we first discuss the mutual information rate more generally and explain how it can be computed via the Gaussian approximation and by using PWS. Then we turn to the two case studies discussed above.

6.1. THE MUTUAL INFORMATION RATE

The mutual information between two random variables S and X is defined as

$$I(S,X) = \iint \mathrm{P}(s,x) \ln \frac{\mathrm{P}(s,x)}{\mathrm{P}(s)\,\mathrm{P}(x)}\, ds\, dx\,, \tag{6.1}$$

or, equivalently, using Shannon entropies

$$\begin{aligned} I(S,X) &= H(S) + H(X) - H(S,X) \\ &= H(S) - H(S|X) \\ &= H(X) - H(X|S)\,. \end{aligned} \tag{6.2}$$

In the context of a noisy communication channel, S and X represent the messages at the sending and receiving end, respectively. Then, $I(S,X)$ is the amount of information about S that is communicated when only X is received. If S can be perfectly reconstructed from X, then $I(S,X) = H(S)$. On the contrary, if S and X are independent, $I(S,X) = 0$. The mutual information thus is always non-negative and quantifies the degree of statistical dependence between two random variables.

For systems that continuously transmit information over time, this concept must be extended to trajectories $\boldsymbol{S}_T = \{S(t) \mid t \in [0,T]\}$ and $\boldsymbol{X}_T = \{X(t) \mid t \in [0,T]\}$. The mutual information between trajectories is defined analogously as

$$I(\boldsymbol{S}_T, \boldsymbol{X}_T) = \left\langle \ln \frac{\mathrm{P}(\boldsymbol{s}_T, \boldsymbol{x}_T)}{\mathrm{P}(\boldsymbol{s}_T)\,\mathrm{P}(\boldsymbol{x}_T)} \right\rangle \tag{6.3}$$

where the expected value is taken with respect to the full joint probability of both trajectories. This quantity can be interpreted as the total information that is communicated over the time interval $[0,T]$.

Note that the total amount of information communicated over the time-interval $[0,T]$ is not directly related to the instantaneous mutual information $I(S(t),X(t))$ at any instant $0 \leq t \leq T$. This is because auto-correlations within the input or output sequences reduce the amount of new information transmitted in subsequent measurements. Moreover, information can be encoded in temporal features of the trajectories, which cannot be captured by an instantaneous information measure. Therefore, as previously pointed out [93, 109], the instantaneous mutual information $I(S(t),X(t))$ for any given t does not provide a meaningful measure of information transmission. To correctly quantify the amount of information transmitted per unit time we must consider entire trajectories.

For that reason, the *mutual information rate* is defined via the trajectory mutual information. Let the input and output of a system be given by two continuous-time stochastic processes $\mathcal{S} = \{S(t) \mid t \in \mathbb{R}\}$ and $\mathcal{X} = \{X(t) \mid t \in \mathbb{R}\}$. Then, the mutual information rate between $\mathcal{S}$ and $\mathcal{X}$ is

$$R(\mathcal{S},\mathcal{X}) = \lim_{T\to\infty} \frac{1}{T} I(\boldsymbol{S}_T, \boldsymbol{X}_T)\,, \tag{6.4}$$

and quantifies the amount of information that can reliably be transmitted per unit time. The mutual information rate therefore represents an excellent performance measure for information processing systems.

In summary, the mutual information rate is *the* crucial performance metric for stochastic information processing systems. However, its information-theoretic definition does not provide an obvious scheme for computing it. As a result, various methods have been developed to compute or approximate the mutual information rate.

6.2. GAUSSIAN APPROXIMATION

One way to significantly simplify the computation of the information rate, is to assume that the input and output trajectories obey stationary Gaussian statistics. Under this assumption Eq. (6.3) simplifies to,

$$I(\boldsymbol{S}_T, \boldsymbol{X}_T) = \frac{1}{2} \ln \frac{|\boldsymbol{C}_{ss}||\boldsymbol{C}_{xx}|}{|\boldsymbol{Z}|}\,, \tag{6.5}$$

where $|\boldsymbol{C}_{ss}|$ and $|\boldsymbol{C}_{xx}|$ are the determinants of the covariance matrices of the respective trajectories $\boldsymbol{S}_T$ and $\boldsymbol{X}_T$, and

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{C}_{ss} & \boldsymbol{C}_{sx} \\ \boldsymbol{C}_{xs} & \boldsymbol{C}_{xx} \end{pmatrix} \tag{6.6}$$

is the covariance matrix of their joint distribution.

In the limit that the trajectory length $N = T/\Delta$, with the discretization Δ, becomes infinitely long ($N \to \infty$) and continuous ($\Delta \to 0$), the information rate as defined in Eq. (6.4) can be expressed in terms of the power spectral densities, or power spectra, of the processes $\mathcal{S}$ and $\mathcal{X}$ [36, 37]:

$$R(\mathcal{S}, \mathcal{X}) = -\frac{1}{4\pi} \int_{-\infty}^{\infty} d\omega \ln\left(1 - \frac{|S_{sx}(\omega)|^2}{S_{ss}(\omega) S_{xx}(\omega)}\right). \tag{6.7}$$

Here, $S_{ss}(\omega)$ and $S_{xx}(\omega)$ respectively are the power spectra of trajectories generated by $\mathcal{S}$ and $\mathcal{X}$, and $S_{sx}(\omega)$ is their cross-spectrum. The fraction $|S_{sx}(\omega)|^2/(S_{ss}(\omega)S_{xx}(\omega))$ is known as the coherence, describing the distribution of power transfer between $\mathcal{S}$ and $\mathcal{X}$ over the frequency ω.

For systems that are neither Gaussian nor linear, there are two ways to still obtain an approximate Gaussian information rate. The first is to directly measure two-point correlation functions from data or simulations, and use these to retrieve the power spectra in Eq. (6.7). The second is to use Van Kampen's linear noise approximation (LNA) and approximate the dynamics of the system to first order around a fixed point [30], see also Appendix 6.A. In this chapter, we will analyze both of these methods.

6.3. PATH WEIGHT SAMPLING FOR DIFFUSIVE SYSTEMS

To evaluate the accuracy of the Gaussian information rate for non-Gaussian systems, an exact method for determining the true information rate is required. Recently, a method called Path Weight Sampling (PWS) was developed, which computes the exact mutual information rate using Monte Carlo techniques without relying on approximations [35].

In Ref. [35], PWS was introduced as a computational framework for calculating the mutual information rate in systems governed by master equations. Master equations provide an exact stochastic description of continuous-time processes with discrete state-spaces, commonly used in models ranging from biochemical signaling networks to population dynamics. However, many systems are not described by discrete state spaces and instead require a stochastic description based on diffusion processes or other stochastic models. Fortunately, PWS is not restricted to systems described by master equations and can be extended to a variety of stochastic models.

In general, PWS can be applied to any system that meets the following conditions: (i) sampling from the input distribution $\mathrm{P}(\boldsymbol{s}_T)$ is straightforward, (ii) sampling from the conditional output distribution $\mathrm{P}(\boldsymbol{x}_T \mid \boldsymbol{s}_T)$ is straightforward, and (iii) the conditional probability density $\mathrm{P}(\boldsymbol{x}_T \mid \boldsymbol{s}_T)$, referred to as the path weight, can be evaluated efficiently. For any stochastic model that satisfies these three criteria, the PWS computation proceeds similarly to systems governed by master equations.

Briefly, PWS computes the trajectory Mutual Information using a Monte Carlo estimate of Eq. (6.3)

$$\frac{\sum_{i=1}^{N}\left[\ln \mathrm{P}\left(\boldsymbol{x}_T^i \,\middle|\, \boldsymbol{s}_T^i\right) - \ln \mathrm{P}\left(\boldsymbol{x}_T^i\right)\right]}{N} \tag{6.8}$$

where $\boldsymbol{s}_T^1, \dots, \boldsymbol{s}_T^N$ are independently drawn from $\mathrm{P}(\boldsymbol{s}_T)$, and each $\boldsymbol{x}_T^i$ is drawn from $\mathrm{P}(\boldsymbol{x}_T \mid \boldsymbol{s}_T^i)$. As $N \to \infty$, this expression converges to the mutual information $I(\boldsymbol{S}_T, \boldsymbol{X}_T)$. In Eq. (6.8), the conditional probability $\mathrm{P}(\boldsymbol{x}_T \mid \boldsymbol{s}_T)$ can be evaluated directly (per criterion iii), but the marginal probability $\mathrm{P}(\boldsymbol{x}_T)$ has to be computed separately for each output trajectory $\boldsymbol{x}_T^i$. Typically, this has to be done numerically via marginalization, i.e., by computing the path integral

$$\mathrm{P}(\boldsymbol{x}_T) = \int d\boldsymbol{s}_T \; \mathrm{P}(\boldsymbol{s}_T)\, \mathrm{P}(\boldsymbol{x}_T \mid \boldsymbol{s}_T) \tag{6.9}$$

using Monte Carlo integration. Evaluating the marginalization integral efficiently is essential for computing the mutual information using PWS and discussed in detail in [35]. In summary, PWS is a generic framework that can be used beyond systems defined by a master equation as long as a suitable generative model satisfying the three conditions above is available.

For this study, we extended PWS to compute the mutual information rate for systems with diffusive dynamics, described by Langevin equations. For such systems, the aforementioned conditions are inherently fulfilled and PWS can be applied. Specifically, in a Langevin system, both the input $S(t)$ and the output $X(t)$ are stochastic processes given by the solution to a stochastic differential equation (SDE). Using stochastic integration schemes like the Euler-Mayurama method, we can straightforwardly generate realizations $s(t)$ and $x(t)$ from the corresponding stochastic process. These realizations are naturally time-discretized with the integration time step Δt. For a time-discretized trajectory $\boldsymbol{x}_{1:n} = (x_1, \dots, x_n)$, the path weight $\ln \mathrm{P}(\boldsymbol{x}_{1:n} \mid \boldsymbol{s}_{1:n})$ is—up to a Gaussian normalization constant—given by the Onsager-Machlup action [110]

$$\ln \mathrm{P}(\boldsymbol{x}_{1:n} \mid \boldsymbol{s}_{1:n}) = -\sum_{i=1}^{n-1} \frac{1}{2\Delta t}\left(\frac{\Delta x_i - v_i \Delta t}{\sigma(x_i)}\right)^2 + \text{cst.} \tag{6.10}$$

where we used $\Delta x_i = x_{i+1} - x_i$, and $v_i = f(x_i, s_i)$ is the deterministic drift, and $\sigma(x_i)$ represents the white noise amplitude. This expression captures the likelihood of a particular trajectory, given the stochastic dynamics of the system, and serves as the path weight in the PWS computation.

6.4. CASE STUDIES

To investigate the conditions under which the Gaussian approximation deviates from the exact mutual information rate, we conducted two case studies. In both studies we compare the Gaussian approximation against the exact mutual information rate, computed via PWS. In the first case study we focus on a discrete linear system which is inspired by minimal motifs of cellular signaling.

DISCRETE REACTION SYSTEM

We consider a simple linear reaction system of two species, S and X, whose dynamics are governed by 4 reactions

$$\varnothing \underset{\lambda}{\overset{\kappa}{\rightleftharpoons}} \mathrm{S} \tag{6.11}$$

$$\mathrm{S} \xrightarrow{\rho} \mathrm{S}+\mathrm{X} \tag{6.12}$$

$$\mathrm{X} \xrightarrow{\mu} \varnothing\,. \tag{6.13}$$

The reaction system is linear because each reaction has at most one reactant. The trajectories of S and X are correlated because the production rate of X depends on the copy number of S, and therefore information is transferred from S to X. This set of reactions can be interpreted as a simple motif for gene expression where S is a transcription factor and X represents the expressed protein. In steady state, the mean copy numbers are given by $\bar{s} = \kappa\lambda^{-1}$ and $\bar{x} = \bar{s}\rho\mu^{-1}$.

The exact stochastic dynamics of this reaction system can be expressed by the chemical master equation [30]. This equation describes the time-evolution of the probability distribution over the possible copy numbers of species S and X, capturing the noise from the discrete chemical reaction events. From this description we can obtain the mutual information rate from S to X without approximations using PWS [35].

While the chemical master equation is an exact representation of the reaction system, for large copy numbers the stochastic dynamics are well-approximated by a linearized model around the steady state. The resulting Langevin equations can be systematically derived from the master equation using the LNA which yields

$$\dot{s}(t) = \kappa - \lambda s(t) + \eta_s(t) \tag{6.14}$$

$$\dot{x}(t) = \rho s(t) - \mu x(t) + \eta_x(t) \tag{6.15}$$

where s and x are continuous variables representing the copy numbers of S and X, and η_s, η_x are independent delta-correlated white noise terms with $\langle\eta_s^2\rangle = 2\lambda\bar{s}$ and $\langle\eta_x^2\rangle = 2\mu\bar{x}$, see Section 6.A.

The Gaussian approximation of the mutual information rate is derived from the LNA description. Using this framework, Tostevin and ten Wolde [36] found an analytical expression for the mutual information rate of the motif in units of nats s^{-1}:

$$R_{\text{Gaussian}} = \frac{\lambda}{2}\left(\sqrt{1+\frac{\rho}{\lambda}}-1\right). \tag{6.16}$$

More recently, Moor and Zechner [102] have derived a different approximation for the mutual information rate of this reaction system by approximating the relevant filtering equation. This approach explicitly differentiates the contributions of individual reactions to the noise amplitude of each component, while the LNA lumps their contributions together. As we will discuss in more detail below, accounting for the noise from each reaction separately better captures the information transmitted in discrete systems, making this “discrete approximation” more accurate than the Gaussian approximation for this case study. Nevertheless, the result is still based on an approximation that is only

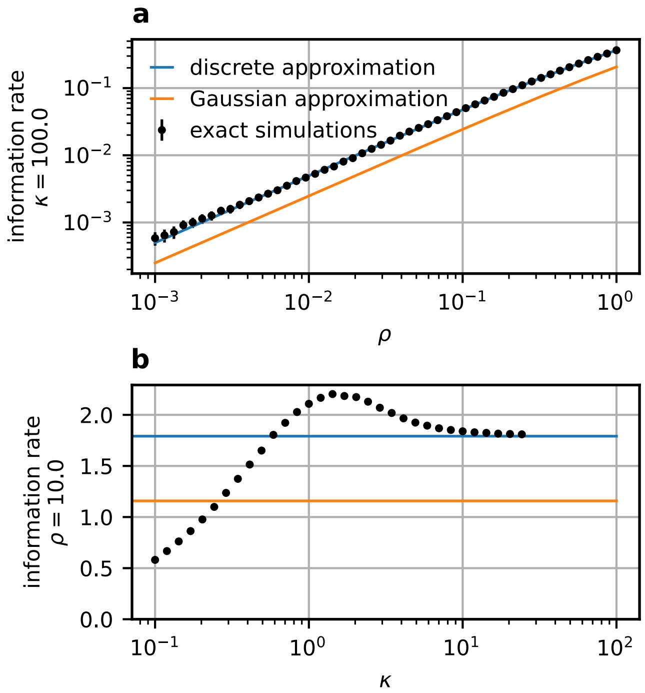


Figure 6.1: The mutual information rate of a simple linear reaction system defined by Eqs. 6.11-6.13. The black dots show the exact information rate, computed with PWS. We compare both the Gaussian approximation of Tostevin and ten Wolde [36] and the discrete approximation of Moor and Zechner [102] against the exact result. In panel (a), we use parameters $\kappa = 100$, $\lambda = 1$, $\mu = 1$ while varying ρ. The mean output copy number is directly proportional to ρ with proportionality factor $\kappa\lambda^{-1} = 100$. In panel (b), we fix $\rho = 10$, $\lambda = 1$, $\mu = 1$, and systematically vary κ. As a consequence, we vary the mean input copy number $\bar{s} = \kappa\lambda^{-1}$, and simultaneously also the mean output copy number $\bar{x} = \bar{s}\rho\mu^{-1} = 10\bar{s}$.

valid for large copy numbers. The expression for the mutual information rate in the discrete approximation is remarkably similar to the expression obtained using the Gaussian framework:

$$R_{\text{discrete}} = \frac{\lambda}{2}\left(\sqrt{1+2\,\frac{\rho}{\lambda}} - 1\right) \tag{6.17}$$

Note that this equation only differs from Eq. (6.16) by the additional factor 2 inside the square root.

The natural—but incorrect—expectation is that for large copy numbers both approximations converge to the true mutual information rate. However, Eqs. (6.16) and (6.17) already reveal that the two approximations do not converge. Indeed, previous work shows that even in the limit of infinite copy numbers, the Gaussian approximation only yields a lower bound, which is not tight [35, 102].

We compare both approximations against exact PWS simulations for different parameters. In Fig. 6.1a, we vary the mean copy number of the readout, $\bar{x}$, by varying its synthesis rate ρ and compute the mutual information rate using both approximations as well as PWS, while keeping the input copy number constant at $\bar{s} = 100$. We observe that the Gaussian approximation via the LNA [Eq. (6.16)] consistently underestimates the mutual information rate. This confirms that even when $\bar{s}$ and $\bar{x}$ are large, the Gaussian approximation only yields a lower bound to the information rate of the discrete linear system. In contrast, the discrete approximation [Eq. (6.17)] coincides with the true mutual information rate obtained from PWS simulations over all output copy numbers $\bar{x}$, even for $\bar{x} \ll 1$.

In Fig. 6.1b, instead of varying the copy number of the output, we vary the copy number of the input by varying the production rate κ. Note that both the Gaussian approximation and the discrete approximation are independent of κ. Yet, we observe that the true mutual information rate is not. For sufficiently large input copy numbers the discrete approximation coincides with the true information rate while the Gaussian information rate remains only a lower bound. Thus, the discrete approximation is highly accurate for $\bar{s} \geq 10$. For small κ where $\bar{s} < 10$, we find that the mutual information rate deviates from both the LNA as well as the discrete approximation. Surprisingly, we find an optimal value of κ for which the mutual information rate is maximized and exceeds both approximations. This implies that at low input copy numbers, the system is able to extract additional information from the discrete input trajectories, which is not accounted for by either of the approximations.

In all cases, we found that the Gaussian approximation deviates significantly from the true information rate for this discrete system. Seemingly paradoxically, the Gaussian approximation based on the LNA does not converge to the true information rate at high copy numbers, even though the LNA approximates the stochastic dynamics extremely well in this regime. In contrast, the discrete approximation from Moor and Zechner [102] does not suffer from this issue. It has been shown that, generally, the Gaussian approximation is a lower bound on the discrete approximation, prompting the question which features in the trajectories are not captured by the Gaussian approximation.

The root cause for the deviations of the Gaussian approximation lies in how the LNA approximates the chemical master equation. While the true dynamics from the chemical master equation give rise to discrete sample paths, i.e., piece-wise constant trajectories connected by instantaneous discontinuous jumps, the LNA approximation yields continuous stochastic trajectories. Our results demonstrate that a discrete sample path of X carries more information about S than the corresponding continuous sample path $x(t)$ would carry about $s(t)$ in the LNA. This is ultimately due to the fact that in the discrete system, each reaction event is unambiguously recorded in the X trajectories and thus different reactions modifying the same species can be distinguished. However, in the continuous LNA description, all reactions that modify X contribute to the noise term in $x(t)$ but their contributions cannot be distinguished from one another, given an observed $x(t)$-trajectory. Specifically, note that for the motif studied here, the decay reaction of the output $X \to \emptyset$ does not carry information on the input fluctuations since its propensity is independent of the input. Yet, it contributes to the overall fluctuations in the output. The Gaussian approximation only considers the total fluctuations in the output, while the discrete approximation correctly distinguishes between the fluctuations induced from production events and decay events. Therefore, the Gaussian approximation consistently underestimates the true information transmission, whereas the discrete approximation does not suffer from this systematic error. This subtle point is reflected in the difference between Eqs. (6.16) and (6.17) and is illustrated in Fig. 6.1.

NONLINEAR CONTINUOUS SYSTEM

Next, we study a nonlinear variant of the reaction system above. In contrast to the previous case study, we deliberately avoid using discrete dynamics, as we already observed that the Gaussian approximation is generally inaccurate in such systems. Instead, we focus solely on continuous Langevin dynamics to explore how an explicitly nonlinear input-output mapping affects the accuracy of the inherently linear Gaussian approximation. We hypothesize that the accuracy of the Gaussian approximation will deteriorate as the degree of nonlinearity increases. To test this hypothesis, we analyze a simple Langevin system with adjustable nonlinearity.

The system is defined by two coupled Langevin equations, one that describes the input, and one that describes the output. The stochastic dynamics of the input $s(t)$ are given by Eq. (6.14). The output dynamics of $x(t)$ are given by

$$\dot{x}(t) = \rho a(s) - \mu x(t) + \eta_x(t) \tag{6.18}$$

with the Hill function

$$a(s) = \begin{cases} \frac{s^n}{K^n+s^n} & \text{if } s \geq 0, \\ 0 & \text{if } s < 0. \end{cases} \tag{6.19}$$

This function serves as a tuneable non-linearity with the Hill coefficient n. As $n \to 0$, the Hill function approaches a shallow and more linear mapping, while for large n, it becomes sigmoidal and highly non-linear. As $n \to \infty$, $a(s)$ approaches the unit step function centered at $s = K$. The so-called static input-output relation specifies the mean output $\bar{x}(s)$ for a static input signal s and is given by $\bar{x}(s) = \rho a(s)/\mu$. The static gain of this system is then defined as the slope of this relation at $s = \bar{s}$, i.e.,

$$g = \frac{\partial \bar{x}(s)}{\partial s}\bigg|_{\bar{s}} = \frac{n[1 - a(\bar{s})]\bar{x}}{\bar{s}}, \tag{6.20}$$

as derived in Section 6.A. Importantly, the gain of the system is directly proportional to the Hill coefficient n, i.e., the gain is directly coupled to the degree of nonlinearity.

Figure 6.2 shows how, on average, the output $x(t)$ at a given time depends on the input $s(t)$ at that same time (solid colored curves). This is the so-called dynamical input-output relation of a system [15]. The solid black curve in Fig. 6.2 represents the static input-output relation. While the static input-output relation is set by the instantaneous function $a(s)$, the dynamical input-output relation depends not only on this function, but also on the timescale of the response $\tau_x = \mu^{-1}$. The reason is that as the output responds more slowly to the input, the temporal input fluctuations are averaged out more. Therefore, the response of the output becomes more shallow for increasing τ_x. Moreover, slower systems with more shallow responses seem increasingly linear (Fig. 6.2).

When using the Langevin extension of PWS we can directly compute the mutual information rate of this nonlinear model, while the Gaussian approximation can only be

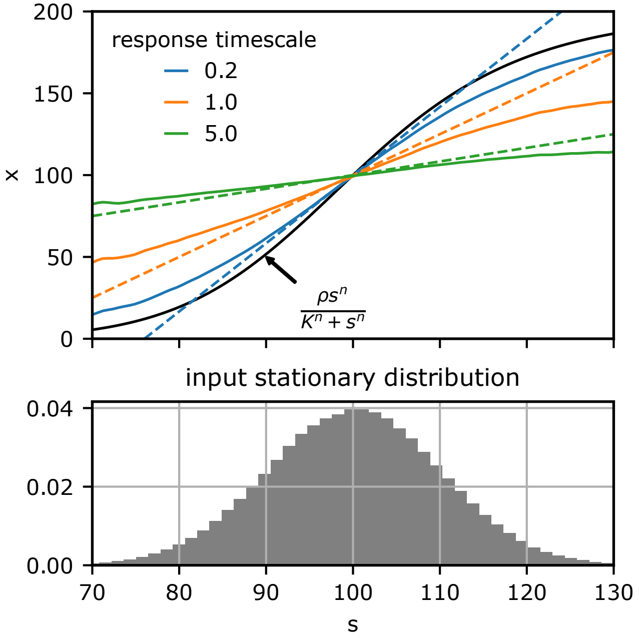


Figure 6.2: The dynamical input output relationship of the non-linear system. The upper panel shows the static input-output relationship (solid black line) as well as the dynamical input output relationship (solid colored lines), i.e., the effective mapping from input to output $s(t) \mapsto x(t)$ for different response timescales $\tau_x = \mu^{-1}$. The static gain g is defined as the slope $\partial_s \bar{x}$ of the static input-output relation at $s = \bar{s}$ and is proportional to the Hill coefficient n. The displayed dynamical input-output mappings were obtained for a (static) gain $g = 5$ [Eq. (6.20)] from simulations of the steady state by sorting the inputs $s(t)$ into bins (bin width 0.5) and computing the average output $x(t)$ corresponding to each bin. Using the linear noise approximation we obtain linear input output mappings with a dynamic gain $\tilde{g} = g/(1+\tau_x)$ (see Appendix 6.A) which are displayed as dashed lines. We observe that the linear mapping approximates the dynamical input output relation well for $s \approx \bar{s} = 100$ but cannot capture the nonlinear saturation effect. The lower panel shows the stationary distribution of $s(t)$.

applied to linear systems. Therefore, to obtain the mutual information rate in the Gaussian approximation, we have to linearize the system. There are two approaches for linearizing the stochastic dynamics of this nonlinear system which result in different information estimates.

The first approach is to linearize Eq. (6.18) analytically via the LNA as shown in Appendix 6.A. Within this approach we can obtain an analytical expression for the information rate [see Eq. (6.41)],

$$R_{\text{LNA}} = \frac{\lambda}{2}\left(\sqrt{1 + g^2 \frac{\bar{s}\mu}{\bar{x}\lambda}} - 1\right), \tag{6.21}$$

This LNA based approach also yields a linearized dynamic input-output relation, shown as dashed lines in Fig. 6.2.

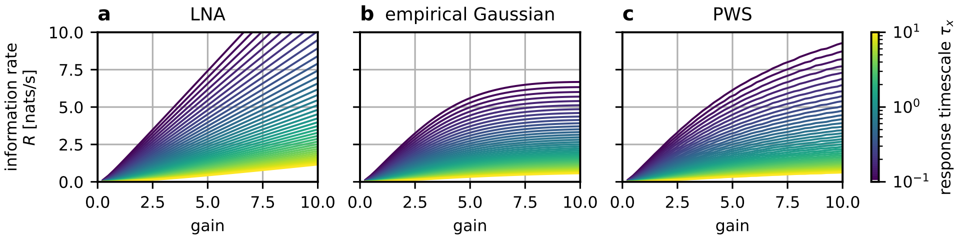


Figure 6.3: The information rate of a non-linear system as a function of its gain over a range of response timescales. We vary the gain g by varying the Hill coefficient n, see Eq. (6.20). A short response timescale corresponds to a fast system (purple) while a long response timescale corresponds to a slow system (yellow). The information rate was computed in three different ways: (a) via the Gaussian approximation using the LNA to estimate the required power spectra; (b) via the Gaussian approximation using simulations to numerically estimate the required power spectra; (c) exactly via PWS.

We observe that the linearized input-output relation closely matches the slope of the true nonlinear dynamical input-output relation at $s \approx \bar{s} = 100$, but overall it does not correspond to a (least-squares) linear fit of the nonlinear dynamical input-output relation. For all values of s, the linearized input-output relation has a slope greater than or equal to the slope of the dynamical input-output relation. Empirically, the LNA thus seems to over-estimate the dynamic gain of the system. The reason may be that the LNA approximates the static input-relation (Fig. 6.2 black curve), and estimates the linearized dynamical input-output relation based on this static approximation only.

The second "empirical Gaussian" approach to linearize the nonlinear system potentially avoids these issues. In this approach, we first numerically generate trajectories from the stochastic Eqs. (6.14) and (6.18) and use digital signal processing techniques to estimate the mutual information rate from the trajectories. We numerically estimate the (cross-)power spectra of input and response using Welch's method [111, see Ch. 11]. From the estimated spectral densities $\hat{S}_{\alpha\beta}(\omega)$ we compute the coherence

$$\hat{\phi}_{sx}(\omega) = \frac{|\hat{S}_{sx}(\omega)|^2}{\hat{S}_{ss}(\omega)\hat{S}_{xx}(\omega)} \tag{6.22}$$

which we use to obtain the Gaussian approximation of the mutual information rate using Eq. (6.7). This estimate assumes that the Fourier components of the signals are statistically independent, i.e., that a hypothetical sinusoidal input signal only leads to output at one unique frequency. While we should thus only expect an accurate result if the statistics of $s(t)$ are Gaussian and the system response is linear [37], the empirical way of obtaining the Gaussian approximation is often the only practical way of computing the mutual information rate directly from experimental data.

The empirical power spectra characterize the linear response of a system, but not in the same way as the LNA. While for linear systems the power spectra obtained via the LNA match the empirical power spectra [112], for a nonlinear system, the empirical power spectra and the coherence can differ from the corresponding LNA calculations. The two linearization approaches are thus not equivalent. We tested the accuracy of

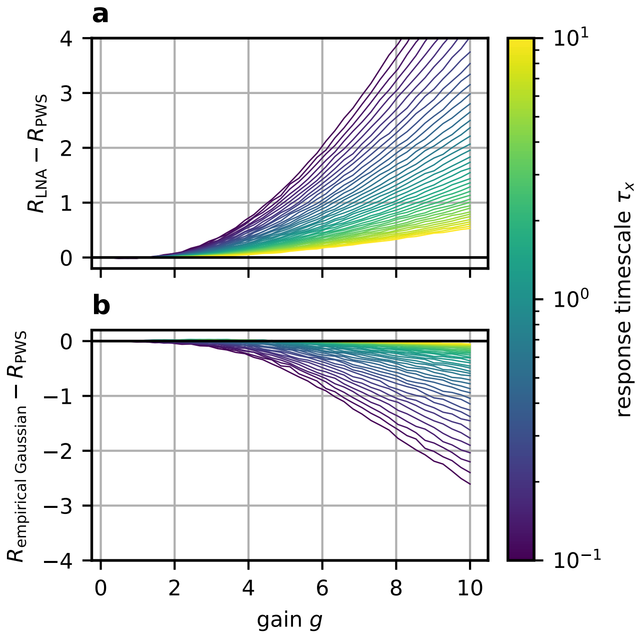


Figure 6.4: Deviation from the exact information rate for the approximate Gaussian information rate computed via LNA (a), and the approximate Gaussian information rate computed via numerical estimates of the power spectra (b). A deviation of 0 implies perfect accuracy. While the absolute deviation of both approximations increases with increasing gain and decreasing response timescale, the LNA based approach consistently overestimates the information rate whereas the empirical approach constitutes a lower bound. Moreover, the empirical approach is more accurate across all parameter values.

the Gaussian mutual information rate estimates using both linearization approaches to elucidate the differences in these approaches.

Figure 6.3 displays the mutual information rate obtained via two linearized approximations as well as the exact PWS result. We vary the gain g and the response time-scale τ_x, both of which significantly affect the shape of the dynamical input-output relationship. As expected, a larger gain or a faster response time lead to an increase in the mutual information rate. At large gain, the information rate naturally saturates as $a(s)$ approaches a step function. The saturation effect is clearly seen in the PWS results, and is found to be even more pronounced in the empirical Gaussian approximation. The LNA-based Gaussian approximation, however, shows no saturation. This highlights that the LNA linearizes the system at the level of the input-output mapping $a(s)$ which results in an approximation that is unaffected by the sigmoidal shape of $a(s)$. In contrast, the empirical approximation is affected by nonlinear saturation effects because it is computed directly from simulated trajectories. We thus see that both approximations yield substantially different results at large gain.

In Fig. 6.4 we compare the absolute deviation between the approximations and the

PWS result. For small gain we see that both approximations are accurate which is not surprising since the nonlinearity is very weak in this regime. Strikingly, for large gain, the LNA-based approximation always overestimates the mutual information rate while the empirical Gaussian approximation always underestimates the rate. In both cases the systematic error decreases as the response timescale becomes slower. This reflects the fact that for slow responders, the dynamic input-output relation is more linear (Fig. 6.2) than for fast responders.

Additionally, we computed the relative deviation, see Fig. 6.5 in Appendix 6.B. We find that in terms of relative error the curves for different response timescales largely overlap. In terms of relative approximation error, the gain, rather than response timescale, is the primary factor affecting the accuracy of the Gaussian approximation.

6.5. Discussion

We investigated the accuracy of the Gaussian approximation for the mutual information rate in two case studies, each highlighting a scenario where the approximation may be inaccurate. We were able to reliably quantify the inaccuracy in each case by computing the "ground truth" mutual information rate for these scenarios using a recently developed exact Monte Carlo technique called PWS [35].

We first considered linear discrete systems, which are relevant in biology due to the discrete nature of biochemical signaling networks. In our example, the Gaussian approximation cannot capture the full information rate, but only yields a lower bound. We show that a different discrete approximation developed by Moor and Zechner [102] is able to correctly estimate the mutual information rate of the network over a wide range of parameters. Since the Gaussian approximation captures the second moments of the discrete system, this finding demonstrates that a discrete system can transmit significantly more information than what would be inferred from its second moments alone. This perhaps surprising fact has been highlighted before in Refs. [35, 102] and it hinges on the use of a discrete reaction-based readout. The increased mutual information rate found for a discrete readout comes from being able to unambiguously distinguish between individual reaction events in the readout's trajectory, which is not possible with a continuous Gaussian readout. The ability to differentiate between production and decay events in the discrete trajectory thus carries a significant amount of information. However, it remains an open question whether biological (or other) signaling systems can effectively harness this additional information encoded in the discrete reaction events. In fact, for systems that cannot distinguish individual reaction events in downstream processing the Gaussian framework might still accurately quantify the "accessible information".

A notable observation in our first case study is the deviation between the discrete approximation of the mutual information rate derived by Moor and Zechner [102] and the exact result obtained using PWS [35] for inputs with low copy number $\bar{s}$. In the discrete approximation, the mutual information rate is independent of the input copy number $\bar{s}$, but the PWS simulations show that at low copy numbers there is an optimal $s^\star \approx 1$ which maximizes the mutual information rate. This surprising finding suggests that the information rate in discrete systems can be increased by reducing the copy number of the input sufficiently, such that it only switches between a few discrete input levels. Notably, in the reverse case—low output copy number $\bar{x}$ but large $\bar{s}$—the discrete approximation always remains accurate. We leave a precise characterization of this finding for future work.

The second example focused on a continuous but nonlinear system, where we demonstrated that the accuracy of the Gaussian approximation depends on the linearization method. Linearizing the underlying system dynamics directly via the LNA leads to an overestimation of the information rate, while estimating the system's correlation functions empirically from data underestimates it. Regardless of the method, the Gaussian approximation is more accurate in terms of absolute deviation of the true information rate when the gain of the system is small or its response is slow compared to the timescale of the input fluctuations.

The result of our second case study—that the empirical Gaussian mutual information rate underestimates the true rate—is consistent with theoretical expectations. As shown by Mitra and Stark [108], and highlighted in [37], an empirical Gaussian estimate of the mutual information between a Gaussian input signal S_G and a non-Gaussian output X provides a lower bound on the channel capacity $C(S,X) = \max_{\mathrm{P}(s)} I(S,X)$ (subject to a power constraint on S). Specifically, they show that $C(S,X) \geq I(S_G;X) \geq I(S_G;X_G)$, where (S_G, X_G) is a jointly Gaussian pair with the same covariance matrix as (S_G, X). For purely Gaussian systems like (S_G, X_G), the mutual information calculated using Eq. (6.7) is exact and equal to the channel capacity. However, for systems that have a Gaussian input but are otherwise non-Gaussian, the mutual information is larger or equal than the corresponding Gaussian model with matching second moments, as evidenced in Fig. 6.4. In general, the empirical Gaussian approximation yields a lower bound on the mutual information of the nonlinear system with a Gaussian input signal, as well as a lower bound on the channel capacity of the nonlinear system[1].

We can distill several concrete recommendations for the computation of the information rate from our analysis. For linear discrete systems, the Gaussian approximation yields a lower bound on the true information rate which may accurately quantify the information available to systems that cannot distinguish individual discrete events. Alternatively, the reaction-based discrete approximation by Moor and Zechner [102] is highly accurate, even when the copy number of the output is extremely small. However, when the copy number of the input becomes small ($\lesssim 10$), both approximations break down and one must use an exact method. Exact methods for obtaining the information rate of any stochastic reaction-based system are PWS [35], or numerical integration of the stochastic filtering equation as shown in [102]. For nonlinear continuous systems with small gain one can safely use the Gaussian approximation, either based on a linearization of the underlying dynamics or on empirically estimated correlation functions. For systems with a high gain but a slow response, one may still use the Gaussian approximation, where an approach based on empirical correlation functions yields the most

[1] Note that this argument does not apply to the Linear Noise Approximation (LNA). The bound specifically requires the Gaussian model to use the covariance of the full, original system. When the system is first linearized using the LNA, the resulting linear model does not retain the same covariance as the original nonlinear system. As a result, the mutual information rate calculated with the LNA is not a lower (nor an upper) bound on the true mutual information rate.

accurate result, and is a rigorous lower bound. Finally, if the system is both highly nonlinear and has a fast response, one must again resort to an exact method like PWS. We hope that our results will guide future research in determining the appropriate method for computing the mutual information rate.

Overall, our results greatly increase the usefulness of the Gaussian approximation for the information rate of non-Gaussian systems. Today, the Gaussian approximation remains the only method that can be applied directly and straightforwardly to experimental data. Here, we have quantified the prerequisites to safely use this approach. Moreover, we elucidate how an appropriate Gaussian approximation always constitutes a lower bound on the true information rate for systems with a sufficiently large input copy number.

APPENDIX

6.A. GAUSSIAN APPROXIMATION

Here we derive the analytical expressions for the Gaussian information rate of the networks considered in the main text. To this end we first discuss the dynamics of the input signal S and its power spectrum. Then, we perform a linear approximation of the dynamics of the readout species X and derive the approximate Gaussian information rate between S and X for the nonlinear network. Finally, we derive the Gaussian information rate of the linear network from our expression of the Gaussian information rate of the nonlinear network.

SIGNAL

The input signal is generated by a birth-death process,

$$\varnothing \underset{\lambda}{\overset{\kappa}{\rightleftharpoons}} \mathrm{S}, \tag{6.23}$$

Its dynamics in Langevin form are,

$$\dot{s} = \kappa - \lambda s(t) + \eta_s(t), \tag{6.24}$$

yielding the steady state signal concentration $\bar{s} = \kappa/\lambda$. The independent Gaussian white noise process $\eta_s(t)$ summarizes all reactions that contribute to fluctuations in S. The strength of the noise term in steady state is

$$\langle \eta_s^2 \rangle = \kappa + \lambda \bar{s} = 2\lambda \bar{s}. \tag{6.25}$$

The power spectral density, or power spectrum, of a stationary process $\mathcal{X}$ is defined as $S_{xx}(\omega) = \lim_{T\to\infty} \frac{1}{T} |\tilde{x}_T(\omega)|^2$, where $\tilde{x}(\omega)$ denotes the Fourier transform of $x(t)$. The power spectrum of a signal obeying Eq. (6.24) is thus given by

$$S_{ss}(\omega) = \frac{\langle \eta_s^2 \rangle}{\omega^2 + \lambda^2} = \frac{2\lambda\bar{s}}{\omega^2 + \lambda^2}. \tag{6.26}$$

LINEAR APPROXIMATION

We now consider the readout X, which is produced via a nonlinear activation function $a(s)$:

$$\begin{aligned} \mathrm{S} &\xrightarrow{\rho\, a(s)} \mathrm{S} + \mathrm{X}, \\ \mathrm{X} &\xrightarrow{\mu} \varnothing. \end{aligned} \tag{6.27}$$

We define the activation level $a(s)$ to be a Hill function,

$$a(s) = \frac{s(t)^n}{K^n + s(t)^n}. \tag{6.28}$$

Such a dependency, in which K sets the concentration of S at which the activation is half-maximal and n sets the steepness, can for example arise from cooperativity between the signal molecules in activating the synthesis of X.

We have for the dynamics of X in Langevin form

$$\dot{x} = \rho\, a(s) - \mu\, x(t) + \eta_x(t), \tag{6.29}$$

with $a(s)$ given by Eq. (6.28). The steady state concentration of X is given by $\bar{x} = \bar{a}\rho/\mu$, where we have defined the steady state activation level $\bar{a} = a(\bar{s})$. It is useful to determine the static gain of the network, which is defined as the change in the steady state of the output upon a change in the steady state of the signal:

$$\begin{aligned} g &= \partial_{\bar{s}}\bar{x} = r/\mu, \\ &= n(1-\bar{a})\bar{x}/\bar{s}, \end{aligned} \tag{6.30}$$

where we have defined the approximate linear activation rate

$$r = n\bar{a}(1-\bar{a})\rho/\bar{s}, \tag{6.31}$$

and the steady state of the activation level is given by

$$\bar{a} = \frac{\bar{s}^n}{K^n + \bar{s}^n}. \tag{6.32}$$

Generally, we assume that $K = \bar{s}$, which entails that in steady state the network is tuned to $\bar{a} = 1/2$.

To compute the Gaussian information rate we approximate the dynamics of X to first nonzero order around $\bar{x}$ via the classical linear noise approximation [30]. Within this approximation the dynamics of the deviation $\delta x(t) = x(t) - \bar{x}$ are,

$$\dot{\delta x} = r\,\delta s(t) - \mu\delta x(t) + \eta_x(t), \tag{6.33}$$

with the synthesis rate r given by Eq. (6.31).

In the linear noise approximation the noise strength is a constant given by the noise strength at steady state,

$$\langle \eta_x^2 \rangle = \rho\bar{a} + \mu\bar{x} = 2\mu\bar{x}. \tag{6.34}$$

INFORMATION RATE

Following Tostevin & Ten Wolde [36, 37], we can express the Gaussian information rate as follows,

$$R(\mathcal{S};\mathcal{X}) = \frac{1}{4\pi}\int_{-\infty}^{\infty} d\omega \log\left(1 + \frac{|K(\omega)|^2}{|N(\omega)|^2} S_{ss}(\omega)\right), \tag{6.35}$$

where $|K(\omega)|^2$ is the frequency dependent gain and $|N(\omega)|^2$ is the frequency dependent noise of the output process $\mathcal{X}$. If the intrinsic noise of the network is not correlated to the process that drives the signal, the power spectrum of the network output obeys the spectral addition rule [31]. In this case the frequency dependent gain and noise can be identified directly from the power spectrum of the output, because it takes the following form:

$$S_{xx}(\omega) = |K(\omega)|^2 S_{ss}(\omega) + |N(\omega)|^2. \tag{6.36}$$

For a species X obeying Eq. (6.33), we have

$$\begin{aligned} |K(\omega)|^2 &= \frac{r^2}{\mu^2+\omega^2}, \\ |N(\omega)|^2 &= \frac{\langle \eta_x^2 \rangle}{\mu^2+\omega^2} = \frac{2\mu\bar{x}}{\mu^2+\omega^2}. \end{aligned} \tag{6.37}$$

The Wiener-Khinchin theorem states that the power spectrum of a stochastic process and its auto-correlation function are a Fourier transform pair. We thus obtain for the variance in the readout, substituting the frequency dependent gain and noise [Eq. (6.37)] and the power spectrum of the signal [Eq. (6.26)] in Eq. (6.36) and taking the inverse Fourier transform at $t = 0$,

$$\sigma_x^2 = g\tilde{g}\sigma_s^2 + \sigma_{x|s}^2 = \frac{g^2\bar{s}}{1+\lambda/\mu} + \bar{x}, \tag{6.38}$$

where the signal variance equals its mean $\sigma_s^2 = \bar{s}$, and the mean readout concentration sets the intrinsic noise $\sigma_{x|s}^2 = \bar{x}$. We further have the static gain g given by Eq. (6.30), and have the dynamic gain

$$\tilde{g} \equiv \frac{\langle \delta x(t)\delta s(t)\rangle}{\sigma_s^2} = \frac{g}{1+\lambda/\mu}, \tag{6.39}$$

which is the slope of the mapping from the time-varying signal value $s(t)$ to the time-varying readout $x(t)$; for Gaussian systems $\langle x(t)|s(t)\rangle = \tilde{g}s(t)$ [15, 37, 86].

To solve the integral in Eq. (6.35) we exploit that

$$\int_{-\infty}^{\infty} d\omega \log\left(\frac{\omega^2+a^2}{\omega^2+b^2}\right) = 2\pi(a-b). \tag{6.40}$$

Substituting the frequency dependent gain and noise given in Eq. (6.37) and the signal power spectrum of Eq. (6.26) in Eq. (6.35) and using Eq. (6.40) we obtain the information rate,

$$\begin{aligned} R(\mathcal{S};\mathcal{X}) &= \frac{\lambda}{2}\left(\sqrt{1+\frac{r^2\langle\eta_s^2\rangle}{\lambda^2\langle\eta_x^2\rangle}}-1\right), \\ &= \frac{\lambda}{2}\left(\sqrt{1+g^2\frac{\bar{s}\mu}{\bar{x}\lambda}}-1\right), \end{aligned} \tag{6.41}$$

where we used the noise strengths given in Eq. (6.25) and Eq. (6.34), we have the static gain g of Eq. (6.30), and the synthesis rate r of Eq. (6.31).

LINEAR NETWORK

To disambiguate differences in the information rate caused by the linear approximation of our nonlinear reaction network on the one hand and the Gaussian approximation of the underlying jump process on the other, we consider the information rate of a linear network. Any difference between the exact information rate and the Gaussian information rate must then be a result of the Gaussian approximation. To this end we use the same input signal [Eq. (6.24)], but we consider a linear activation of the readout, i.e.

$$\begin{aligned} \mathrm{S} &\xrightarrow{\rho} \mathrm{S}+\mathrm{X}, \\ \mathrm{X} &\xrightarrow{\mu} \varnothing \end{aligned} \tag{6.42}$$

such that the Langevin dynamics of X are

$$\dot{x} = \rho s(t) - \mu x(t) + \eta_x(t), \tag{6.43}$$

which yields the steady state concentration $\bar{x} = \bar{s}\rho/\mu$. For this linear readout, the static gain is simply set by the ratio of steady states of the input and the output, $g = \rho/\mu = \bar{x}/\bar{s}$. We can then obtain the information rate of this linear system by substitution of its static gain in Eq. (6.41), which yields

$$R(\mathcal{S};\mathcal{X}) = \frac{\lambda}{2}\left(\sqrt{1+\frac{\rho}{\lambda}} - 1\right). \tag{6.44}$$

This result is identical to that of Tostevin and ten Wolde [36, 37] (motif III).

6.B. Relative deviation of Gaussian approximation for a nonlinear system

Due to the definition of the mutual information, an absolute difference in information maps to a relative difference in the reduction of uncertainty. Therefore, Fig. 6.4 in Section 6.4 focuses on the absolute deviation between the Gaussian approximation and the true mutual information. Nevertheless, the relative deviation of the Gaussian information rate from the true rate can still provide valuable insights, and we discuss it here.

In Fig. 6.5 we compare the relative deviation $R_{\mathrm{Gaussian}}/R_{\mathrm{PWS}}$ between the Gaussian approximation and the exact mutual information computed using PWS. We find that the relative deviation increases as the system gain increases, indicating that the Gaussian approximation also becomes relatively less accurate for larger gains. As already discussed above, the empirical Gaussian method consistently underestimates the true information rate, while the LNA-based approximation overestimates it.

Interestingly, we also observe that for the LNA approximation, at fast timescales the result is slightly more accurate, whereas the empirical Gaussian estimate is more accurate at slow timescales. We initially expected that in both cases slow timescales would yield better agreement with PWS, as the input-output dynamics are more linear for slow timescales, and thus better approximated by the Gaussian model. The fact that this is not the case for the LNA approximation is intriguing, indicating the need for further investigation into the interplay between timescales, system nonlinearity, and the LNA.

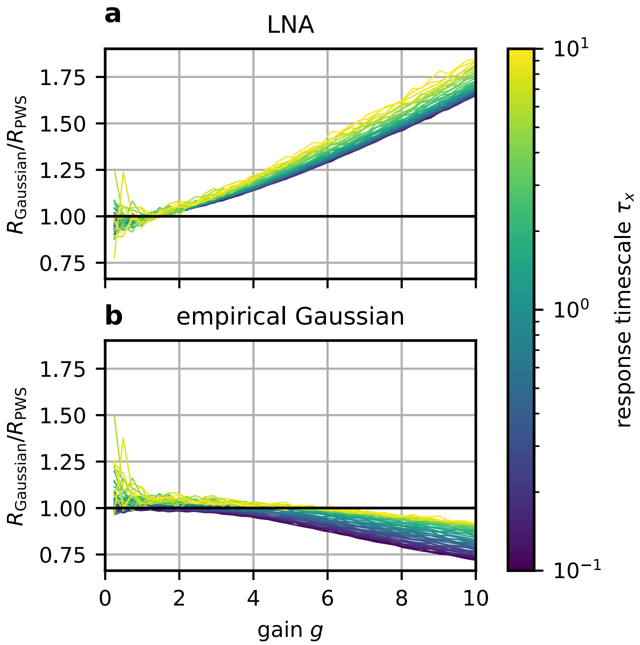


Figure 6.5: Relative deviation from the exact information rate for the approximate Gaussian information rate. The relative deviation was computed for both the LNA-based approximation and the empirical Gaussian approximation using numerically estimated power spectra, across varying system gains and response timescales (see Section 6.4 for details). The curves for different response timescales largely overlap, indicating that the relative approximation error is primarily influenced by system gain rather than response timescale. This highlights that system gain is the dominant factor in determining the accuracy of the Gaussian approximation.

REFERENCES

[1] M. Monti, D. K. Lubensky, and P. R. ten Wolde, *Robustness of Clocks to Input Noise,* Physical Review Letters **121**, 078101 (2018).

[2] W. Pittayakanchit, Z. Lu, J. Chew, M. J. Rust, and A. Murugan, *Biophysical clocks face a trade-off between internal and external noise resistance.* eLife **7** (2018).

[3] E. Kussell and S. Leibler, *Phenotypic diversity, population growth, and information in fluctuating environments,* Science **309**, 2075 (2005).

[4] W. Bialek, *Biophysics: searching for principles*, edited by P. E. A. Inc. (Princeton University Press, Woodstock, Oxfordshire, 2012).

[5] I. Tagkopoulos, Y.-C. Liu, and S. Tavazoie, *Predictive behavior within microbial genetic networks,* Science **320**, 1313 (2008).

[6] A. Mitchell, G. H. Romano, B. Groisman, A. Yona, E. Dekel, M. Kupiec, O. Dahan, and Y. Pilpel, *Adaptive prediction of environmental changes by microorganisms,* Nature **460**, 220 (2009).

[7] N. B. Becker, A. Mugler, and P. R. ten Wolde, *Optimal prediction by cellular signaling networks,* Physical review letters **115**, 258103 (2015).

[8] W. Bialek, R. R. D. R. Van Steveninck, and N. Tishby, *Efficient representation as a design principle for neural coding and computation,* in *2006 IEEE international symposium on information theory* (IEEE, 2006) pp. 659–663.

[9] N. Tishby, F. C. Pereira, and W. Bialek, *The information bottleneck method,* Proceedings of 37th Allerton Conference on communication and computation (1999).

[10] V. Sachdeva, T. Mora, A. M. Walczak, and S. E. Palmer, *Optimal prediction with resource constraints using the information bottleneck,* PLOS Computational Biology **17**, e1008743 (2021).

[11] S. E. Palmer, O. Marre, M. J. Berry, and W. Bialek, *Predictive information in a sensory population,* Proceedings of the National Academy of Sciences of the United States of America **112**, 6908 (2015).

[12] M. Chalk, O. Marre, and G. Tkačik, *Toward a unified theory of efficient, predictive, and sparse coding,* Proceedings of the National Academy of Sciences of the United States of America **115**, 186 (2018).

[13] C. C. Govern and P. R. ten Wolde, *Optimal resource allocation in cellular sensing systems,* Proceedings of the National Academy of Sciences of the United States of America **111**, 17486 LP (2014).

[14] S. B. Laughlin, R. R. De Ruyter van Steveninck, and J. C. Anderson, *The metabolic cost of neural information,* Nature Neuroscience **1**, 36 (1998).

[15] G. Malaguti and P. R. ten Wolde, *Theory for the optimal detection of time-varying signals in cellular sensing systems,* eLife **10**, 1 (2021).

[16] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, *Information Bottleneck for Gaussian Variables,* Journal of Machine Learning Research **6**, 165 (2005).

[17] A. Goldbeter and D. E. Koshland, *An amplified sensitivity arising from covalent modification in biological systems,* Proceedings of the National Academy of Sciences of the United States of America **78**, 6840 (1981).

[18] U. Alon, *An Introduction to Systems Biology: Design Principles of Biological Circuits* (Chapman and Hall/CRC, New York, 2006).

[19] J. E. Segall, S. M. Block, and H. C. Berg, *Temporal comparisons in bacterial chemotaxis.* Proceedings of the National Academy of Sciences of the United States of America **83**, 8987 (1986).

[20] H. H. Mattingly, K. Kamino, B. B. Machta, and T. Emonet, *Escherichia coli chemotaxis is information limited,* Nature Physics 2021 17:12 **17**, 1426 (2021).

[21] R. M. Macnab and D. E. Koshland Jr, *The gradient-sensing mechanism in bacterial chemotaxis,* Proceedings of the National Academy of Sciences **69**, 2509 (1972).

[22] R. Lux and W. Shi, *Chemotaxis-guided movements in bacteria,* Critical Reviews in Oral Biology & Medicine **15**, 207 (2004).

[23] M. D. Baker, P. M. Wolanin, and J. B. Stock, *Systems biology of bacterial chemotaxis,* Current opinion in Microbiology **9**, 187 (2006).

[24] C. V. Rao, G. D. Glekas, and G. W. Ordal, *The three adaptation systems of bacillus subtilis chemotaxis,* Trends in microbiology **16**, 480 (2008).

[25] U. B. Kaupp, J. Solzin, E. Hildebrand, J. E. Brown, A. Helbig, V. Hagen, M. Beyermann, F. Pampaloni, and I. Weyand, *The signal flow and motor response controling chemotaxis of sea urchin sperm,* Nature cell biology **5**, 109 (2003).

[26] B. M. Friedrich and F. Jülicher, *Chemotaxis of sperm cells,* Proceedings of the National Academy of Sciences **104**, 13256 (2007).

[27] M. Abdelgalil, Y. Aboelkassem, and H. Taha, *Sea urchin sperm exploit extremum seeking control to find the egg,* Physical Review E **106**, L062401 (2022).

[28] S. R. Lockery, *The computational worm: spatial orientation and its neuronal basis in c. elegans,* Current opinion in neurobiology **21**, 782 (2011).

[29] G. Malaguti and P. R. ten Wolde, *Receptor time integration via discrete sampling,* Physical Review E **105**, 054406 (2022).

[30] N. Van Kampen, *Stochastic processes in physics and chemistry* (North Holland, Amsterdam, 1992).

[31] S. Tănase-Nicola, P. B. Warren, and P. R. ten Wolde, *Signal detection, modularity, and the correlation between extrinsic and intrinsic noise in biochemical networks,* Physical review letters **97**, 068102 (2006).

[32] E. Ziv, I. Nemenman, and C. H. Wiggins, *Optimal signal processing in small stochastic biochemical networks,* PloS one **2**, e1077 (2007).

[33] W. De Ronde, F. Tostevin, and P. R. ten Wolde, *Effect of feedback on the fidelity of information transmission of time-varying signals,* Phys. Rev. E **82**, 031914 (2010).

[34] T.-L. Wang, B. Kuznets-Speck, J. Broderick, and M. Hinczewski, *The price of a bit: energetic costs and the evolution of cellular signaling,* bioRxiv , 2020.10.06.327700 (2022).

[35] M. Reinhardt, G. Tkačik, and P. R. ten Wolde, *Path weight sampling: Exact monte carlo computation of the mutual information between stochastic trajectories,* Physical Review X **13**, 041017 (2023).

[36] F. Tostevin and P. R. ten Wolde, *Mutual information between input and output trajectories of biochemical networks,* Physical review letters **102**, 218101 (2009).

[37] F. Tostevin and P. R. ten Wolde, *Mutual information in time-varying biochemical systems,* Physical Review E **81**, 061917 (2010).

[38] A. J. Tjalma, V. Galstyan, J. Goedhart, L. Slim, N. B. Becker, and P. R. ten Wolde, *Trade-offs between cost and information in cellular prediction,* Proceedings of the National Academy of Sciences **120**, e2303078120 (2023).

[39] J. Poulton, A. Tjalma, L. Slim, and P. R. t. Wolde, *Predicting a noisy signal: the costs and benefits of time averaging as a noise mitigation strategy,* arXiv preprint arXiv:2307.03006 (2023).

[40] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (John Wiley & Sons, 2006).

[41] C. E. Shannon, *A mathematical theory of communication,* The Bell system technical journal **27**, 379 (1948).

[42] W. Bialek and N. Tishby, *Predictive information,* arXiv preprint arXiv:cond-mat/9902341 (1999).

[43] W. Bialek, I. Nemenman, and N. Tishby, *Predictability, complexity, and learning,* Neural computation **13**, 2409 (2001).

[44] N. Wiener, *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications* (The MIT press, 1949).

[45] M. Vennettilli, S. Saha, U. Roy, and A. Mugler, *Precision of Protein Thermometry,* Physical Review Letters **127**, 098102 (2021).

[46] H. H. Mattingly, K. Kamino, J. Ong, R. Kottou, T. Emonet, and B. B. Machta, *E. coli do not count single molecules,* arXiv preprint arXiv:2407.07264 (2024).

[47] T. E. Ouldridge, C. C. Govern, and P. R. ten Wolde, *Thermodynamics of computational copying in biochemical systems,* Physical Review X **7** (2017).

[48] M. Hinczewski and D. Thirumalai, *Cellular Signaling Networks Function as Generalized Wiener-Kolmogorov Filters to Suppress Noise,* Physical Review X **4**, 3 (2014).

[49] M. Li and G. L. Hazelbauer, *Cellular stoichiometry of the components of the chemotaxis signaling complex,* Journal of Bacteriology **186**, 3687 (2004).

[50] V. Sourjik and H. C. Berg, *Functional interactions between receptors in bacterial chemotaxis,* Nature 2004 428:6981 **428**, 437 (2004).

[51] D. T. Gillespie, *Chemical Langevin equation,* Journal of Chemical Physics **113**, 297 (2000).

[52] J. E. Keymer, R. G. Endres, M. Skoge, Y. Meir, and N. S. Wingreen, *Chemosensing in Escherichia coli: Two regimes of two-state receptors,* Proceedings of the National Academy of Sciences of the United States of America **103**, 1786 (2006).

[53] N. Barkai and S. Leibler, *Robustness in simple biochemical networks,* Nature **387**, 913 (1997).

[54] R. G. Endres and N. S. Wingreen, *Precise adaptation in bacterial chemotaxis through "assistance neighborhoods",* Proceedings of the National Academy of Sciences of the United States of America **103**, 13040 (2006).

[55] G. Lan, P. Sartori, S. Neumann, V. Sourjik, and Y. Tu, *The energy–speed–accuracy trade-off in sensory adaptation,* Nature physics **8**, 422 (2012).

[56] V. Sourjik and H. C. Berg, *Binding of the Escherichia coli response regulator CheY to its target measured in vivo by fluorescence resonance energy transfer,* Proceedings of the National Academy of Sciences of the United States of America **99**, 12669 (2002).

[57] B. A. Mello and Y. Tu, *Effects of adaptation in maintaining high sensitivity over a wide range of backgrounds for Escherichia coli chemotaxis,* Biophysical Journal **92**, 2329 (2007).

[58] Y. Tu, T. S. Shimizu, and H. C. Berg, *Modeling the chemotactic response of Escherichia coli to time-varying stimuli,* Proceedings of the National Academy of Sciences of the United States of America **105**, 14855 (2008).

[59] P. Sartori and Y. Tu, *Free Energy Cost of Reducing Noise while Maintaining a High Sensitivity,* Physical Review Letters **115**, 118102 (2015).

[60] J. R. Maddock and L. Shapiro, *Polar Location of the Chemoreceptor Complex in the Escherichia coli Cell,* Science **259**, 1717 (1993).

[61] T. A. J. Duke and D. Bray, *Heightened sensitivity of a lattice of membrane receptors,* Proceedings of the National Academy of Sciences of the United States of America **96**, 10104 (1999).

[62] T. S. Shimizu, Y. Tu, and H. C. Berg, *A modular gradient-sensing network for chemotaxis in Escherichia coli revealed by responses to time-varying stimuli,* Molecular Systems Biology **6**, 1 (2010).

[63] J. M. Keegstra, K. Kamino, F. Anquez, M. D. Lazova, T. Emonet, and T. S. Shimizu, *Phenotypic diversity and temporal variability in a bacterial signaling network revealed by single-cell FRET.* eLife **6**, 708 (2017).

[64] J. S. Parkinson, G. L. Hazelbauer, and J. J. Falke, *Signaling and sensory adaptation in Escherichia coli chemoreceptors: 2015 update,* Trends in Microbiology Special Issue: Microbial Translocation, **23**, 257 (2015).

[65] J. Monod, J. Wyman, and J. P. Changeux, *On the nature of allosteric transitions: A plausible model,* Journal of Molecular Biology **12**, 88 (1965).

[66] B. A. Mello and Y. Tu, *An allosteric model for heterogeneous receptor complexes: Understanding bacterial chemotaxis responses to multiple stimuli,* Proceedings of the National Academy of Sciences of the United States of America **102**, 17354 (2005).

[67] K. Kamino, J. M. Keegstra, J. Long, T. Emonet, and T. S. Shimizu, *Adaptive tuning of cell sensory diversity without changes in gene expression,* Science Advances **6**, eabc1087 (2020).

[68] T. Emonet and P. Cluzel, *Relationship between cellular response and behavioral variability in bacterial chemotaxis,* Proceedings of the National Academy of Sciences of the United States of America **105**, 3304 (2008).

[69] M. N. Levit, T. W. Grebe, and J. B. Stock, *Organization of the Receptor-Kinase Signaling Array That Regulates Escherichia coli Chemotaxis *,* Journal of Biological Chemistry **277**, 36748 (2002).

[70] C. Walsh, *Posttranslational Modification of Proteins: Expanding Nature's Inventory* (Roberts & Company Publishers, 2006).

[71] N. R. Francis, M. N. Levit, T. R. Shaikh, L. A. Melanson, J. B. Stock, and D. J. DeRosier, *Subunit Organization in a Soluble Complex of Tar, CheW, and CheA by Electron Microscopy,* Journal of Biological Chemistry **277**, 36755 (2002).

[72] V. Sourjik and H. C. Berg, *Receptor sensitivity in bacterial chemotaxis,* Proceedings of the National Academy of Sciences of the United States of America **99**, 123 (2002).

[73] M. Bauer, M. D. Petkova, T. Gregor, E. F. Wieschaus, and W. Bialek, *Trading bits in the readout from a genetic network,* Proceedings of the National Academy of Sciences of the United States of America **118**, e2109011118 (2021).

[74] J. M. Keegstra, F. Avgidis, Y. Mulla, J. S. Parkinson, and T. S. Shimizu, *Near-critical tuning of cooperativity revealed by spontaneous switching in a protein signalling array,* bioRxiv preprint 10.1101/2022.12.04.518992 (2023).

[75] D. Hathcock, Q. Yu, B. A. Mello, D. N. Amin, G. L. Hazelbauer, and Y. Tu, *A nonequilibrium allosteric model for receptor-kinase complexes: The role of energy dissipation in chemotaxis signaling,* Proceedings of the National Academy of Sciences **120**, e2303115120 (2023).

[76] D. Hathcock, Q. Yu, and Y. Tu, *Time-reversal symmetry breaking in the chemosensory array reveals a general mechanism for dissipation-enhanced cooperative sensing,* Nature Communications **15**, 8892 (2024).

[77] D. M. Sherry, I. R. Graf, S. J. Bryant, T. Emonet, and B. M. Machta, *Lattice ultrasensitivity produces large gain in e. coli chemosensing,* bioRxiv , 2024 (2024).

[78] B. A. Mello, L. Shaw, and Y. Tu, *Effects of receptor interaction in bacterial chemotaxis,* Biophysical journal **87**, 1578 (2004).

[79] J. P. Moore, K. Kamino, R. Kottou, T. S. Shimizu, and T. Emonet, *Signal integration and adaptive sensory diversity tuning in escherichia coli chemotaxis,* Cell Systems **15**, 628 (2024).

[80] N. Shiraishi, K. Funo, and K. Saito, *Speed Limit for Classical Stochastic Processes,* Physical Review Letters **121**, 070601 (2018).

[81] S. M. Block, J. E. Segall, and H. C. Berg, *Adaptation kinetics in bacterial chemotaxis,* Journal of bacteriology **154**, 312 (1983).

[82] E. A. Korobkova, T. Emonet, H. Park, and P. Cluzel, *Hidden stochastic nature of a single bacterial motor,* Physical review letters **96**, 058105 (2006).

[83] J. O. Dubuis, G. Tkacik, E. F. Wieschaus, T. Gregor, and W. Bialek, *Positional information, in bits,* Proceedings of the National Academy of Sciences of the United States of America **110**, 16301 (2013).

[84] N. I. Markevich, J. B. Hoek, and B. N. Kholodenko, *Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades,* The Journal of cell biology **164**, 353 (2004).

[85] K. Takahashi, S. Tănase-Nicola, and P. R. Ten Wolde, *Spatio-temporal correlations can drastically change the response of a mapk pathway,* Proceedings of the National Academy of Sciences **107**, 2473 (2010).

[86] A. J. Tjalma and P. R. ten Wolde, *Predicting concentration changes via discrete receptor sampling,* Physical Review Research **6**, 033049 (2024).

[87] H. C. Berg and E. M. Purcell, *Physics of chemoreception,* Biophysical Journal **20**, 193 (1977).

[88] N. Barkai and S. Leibler, *Robustness in simple biochemical networks to transfer and process information.* Nature **387**, 913 (1997).

[89] T. M. Yi, Y. Huang, M. I. Simon, and J. Doyle, *Robust perfect adaptation in bacterial chemotaxis through integral feedback control,* Proceedings of the National Academy of Sciences of the United States of America **97**, 4649 (2000).

[90] B. A. Mello and Y. Tu, *Perfect and near-perfect adaptation in a model of bacterial chemotaxis,* Biophysical Journal **84**, 2943 (2003).

[91] A. Tjalma, *Predicting concentration changes via discrete sampling,* (2024), 10.5281/zenodo.11198862.

[92] A. Briegel, X. Li, A. M. Bilwes, K. T. Hughes, G. J. Jensen, and B. R. Crane, *Bacterial chemoreceptor arrays are hexagonally packed trimers of receptor dimers networked by rings of kinase and coupling proteins,* Proceedings of the National Academy of Sciences **109**, 3766 (2012).

[93] M. Meijers, S. Ito, and P. R. ten Wolde, *Behavior of information flow near criticality,* Physical Review E **103**, L010102 (2021).

[94] J. F. Staropoli and U. Alon, *Computerized analysis of chemotaxis at different stages of bacterial growth,* Biophysical Journal **78**, 513 (2000).

[95] L. Turner, W. S. Ryu, and H. C. Berg, *Real-time imaging of fluorescent flagellar filaments,* Journal of bacteriology **182**, 2793 (2000).

[96] N. C. Darnton, L. Turner, S. Rojevsky, and H. C. Berg, *On torque and tumbling in swimming escherichia coli,* Journal of bacteriology **189**, 1756 (2007).

[97] N. W. Frankel, W. Pontius, Y. S. Dufour, J. Long, L. Hernandez-Nunez, and T. Emonet, *Adaptability of non-genetic diversity in bacterial chemotaxis,* Elife **3**, e03526 (2014).

[98] Y. S. Dufour, X. Fu, L. Hernandez-Nunez, and T. Emonet, *Limits of feedback control in bacterial chemotaxis,* PLoS computational biology **10**, e1003694 (2014).

[99] D. R. Brumley, F. Carrara, A. M. Hein, Y. Yawata, S. A. Levin, and R. Stocker, *Bacteria push the limits of chemotactic precision to navigate dynamic chemical gradients,* Proceedings of the National Academy of Sciences **116**, 10792 (2019).

[100] J. M. Keegstra, F. Carrara, and R. Stocker, *The ecological roles of bacterial chemotaxis,* Nature Reviews Microbiology **20**, 491 (2022).

[101] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell,* 5th ed. (Garland Science, 2008) pp. 917-919.

[102] A.-L. Moor and C. Zechner, *Dynamic information transfer in stochastic biochemical networks,* Physical Review Research **5**, 013032 (2023).

[103] L. Hahn, A. M. Walczak, and T. Mora, *Dynamical Information Synergy in Biochemical Signaling Networks,* Physical Review Letters **131**, 128401 (2023).

[104] M. W. Covert, T. H. Leung, J. E. Gaston, and D. Baltimore, *Achieving stability of lipopolysaccharide-induced nf-κb activation,* Science **309**, 1854 (2005).

[105] S. Strong, R. D. R. Van Steveninck, W. Bialek, R. Koberle, *et al.*, *On the application of information theory to neural spike trains,* in *Pac Symp Biocomput,* Vol. 1998 (1998) pp. 621–632.

[106] R. Cheong, A. Rhee, C. J. Wang, I. Nemenman, and A. Levchenko, *Information transduction capacity of noisy biochemical signaling networks,* science **334**, 354 (2011).

[107] G. Tkačik, C. G. Callan, Jr., and W. Bialek, *Information flow and optimization in transcriptional regulation,* Proceedings of the National Academy of Sciences **105**, 12265 (2008), 0705.0313 .

[108] P. P. Mitra and J. B. Stark, *Nonlinear limits to the information capacity of optical fibre communications,* Nature **411**, 1027 (2001).

[109] R. Fan and A. Hilfinger, *Characterizing the nonmonotonic behavior of mutual information along biochemical reaction cascades,* Physical Review E **110**, 034309 (2024).

[110] L. Onsager and S. Machlup, *Fluctuations and Irreversible Processes,* Physical Review **91**, 1505 (1953).

[111] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing* (Prentice-Hall, 1975).

[112] P. B. Warren, S. Tânase-Nicola, and P. R. Ten Wolde, *Exact results for noise power spectra in linear biochemical reaction networks,* Journal of Chemical Physics **125**, 144904 (2006).

SUMMARY

A remarkable feat of life is the ability of organisms to predict future changes. While this complex task is often exclusively associated with higher organisms, even bacteria demonstrate predictive capabilities. These single-celled organisms often live in strongly varying environments, and any cellular response takes time to mount. Anticipating changes in the environment allows cells to mount a response ahead of time, potentially providing an evolutionary advantage. The extent to which a prediction enhances fitness depends on the accuracy of the prediction. However, what determines the accuracy with which single-celled organisms can predict the future remains unclear.

Any information about the future must be obtained from the current or past environment. Consequently, the amount of predictive information a cell can obtain is inherently limited by the amount of information it collects from the past. In Chapter 2, we investigate this fundamental information bound for two different classes of input signals. Our analysis reveals that, for a Markovian input signal, the optimal system should simply copy the most recent signal value into its output in order to reach the information bound. Furthermore, we show that in the presence of high-frequency noise in the signal, the optimal system should not only measure the most recent signal value but also take into account earlier signal values to average out the noise. For a simple class of non-Markovian signals, we find that the optimal system must base its output on both the most recent signal value and its derivative. If the objective is to predict multiple distinct signal properties, the input should also be mapped onto multiple output components, given sufficiently large past information.

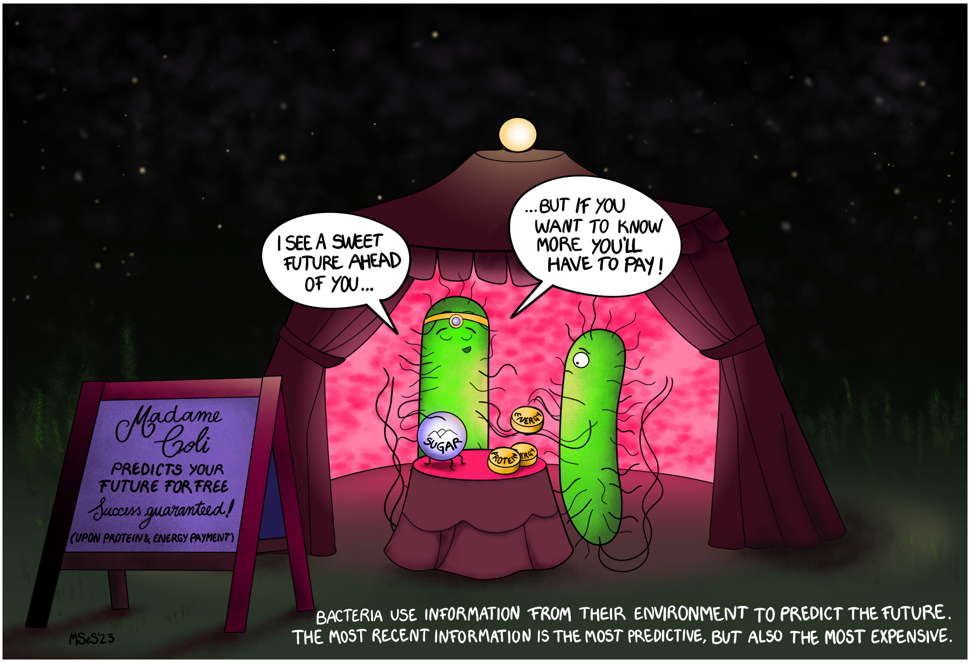
What signal features a biochemical network can detect depends on its topology. In Chapter 3, we show that the canonical push-pull motif can reach the information bound for a Markovian input signal. Yet, how much information the network can extract from the past depends on the availability of physical resources devoted to this task. Considering that the proteins constituting the network must be continually synthesized to prevent dilution by growth and that driving the push-pull motif requires chemical power, we define a resource cost function for the network. Constraining this cost function reveals that, while theoretically possible, reaching the information bound is prohibitively costly. The reason is two-fold: firstly, the chemical power required to drive the network diverges as the copying speed increases; secondly, even in a regime where the cost of driving the network is negligible, the cell can increase both the past and the predictive information by sampling receptors states over time. By increasing both predictive and past information simultaneously, this time-averaging strategy moves the system away from the information bound. This observation highlights that not all bits of past information are equally predictive, nor costly. For the push-pull network, and perhaps more generally, the most recent past information is the most predictive, but also the most costly to obtain.

In Chapter 4 we turn to an important and well-studied biological scenario: chemo-

taxis in shallow concentration gradients. Our analysis shows that the *Escherichia coli* chemotaxis network can reach the information bound for predicting future concentration changes of a signal that follows experimentally measured statistics. However, we again find that reaching the information bound is exceedingly costly. To reach the bound, the network must take an instantaneous derivative. But under constrained resource availability, the gain of temporal derivative-taking networks like that of *E. coli* diminishes with the timescale over which the derivative is measured, known as the adaptation time. Therefore, to lift the signal above the inevitable biochemical noise when resources are limited, the cell must increase the adaptation time, which increases both the past and the predictive information and again moves the system away from the information bound. Computing the past and predictive information of the *E. coli* chemotaxis network directly from experimental data reveals that, given limited resource availability, *E. coli* predicts optimally in the most shallow concentration gradients, which are also the hardest to sense. This finding suggests that it is most important for *E. coli* to use its resources efficiently for prediction in shallow gradients.

E. coli is not the only organism that measures temporal derivatives. Indeed, computing such derivatives by comparing the current signal to that in the recent past, over the adaptation time, appears to be a common navigation strategy. However, it remains unclear what precisely limits the accuracy of such measurements, and sets the optimal adaptation time. In Chapter 5, we extend the previously established sampling framework for the purpose of prediction and use it to study how input statistics and resource availability determine the optimal design of the *E. coli* chemotaxis network. We find that an optimal adaptation time arises from a trade-off between two distinct types of error in the measurement of the derivative: the sampling error, a statistical measurement error caused by the stochastic nature of the biochemical network, and the dynamical error, a systematic error caused by uninformative variations in the input over the duration of the measurement. Increased resource availability reduces the sampling error, allowing for a shorter adaptation time, which in turn reduces the dynamical error. Investigating how the optimal adaptation time depends on the steepness of chemical gradients that the cell navigates, we find that the optimal adaptation time scales inversely with the gradient steepness, as steeper gradients lift the signal above the noise and reduce the sampling error.

Throughout all chapters we have considered input signals with Gaussian statistics, and studied biochemical signaling networks in the linear noise approximation. Previous work has shown that this approach provides a good approximation of the instantaneous mutual information between the input and output of non-Gaussian systems. However, information is often encoded in the dynamics of the full input and output trajectories. This information can be quantified by the mutual information rate. A Gaussian approximation has been established to approximate the mutual information rate, but computing it exactly has proven notoriously difficult due to the high dimensionality of the trajectory space. Therefore, it is unclear how accurate the Gaussian information rate is for non-Gaussian systems. In Chapter 6, we exploit a recent method that enables exact computation of the mutual information to study the accuracy of the Gaussian approximation. We find that for discrete linear reaction networks, the Gaussian approximation only yields a lower bound on the true information rate, because it cannot distinguish between separate reactions, e.g. synthesis and decay of the same molecule. For continuous nonlinear reaction networks, the Gaussian approximation is accurate if the gain of the network is sufficiently small, such that the input-output relation of the network is approximately linear over the typical range of input concentrations. Moreover, if the response of the network is slow relative to the timescale of the input fluctuations, the dynamic input-output relation of the network also becomes more linear. For such systems the Gaussian approximation can be accurate if the two-point correlation functions of the nonlinear network are estimated directly from data, and not via a linear approximation.


I SEE A SWEET FUTURE AHEAD OF YOU...
...BUT IF YOU WANT TO KNOW MORE YOU'LL HAVE TO PAY!
Madame Coli
PREDICTS YOUR FUTURE FOR FREE
Success guaranteed!
(UPON PROTEIN & ENERGY PAYMENT)
SUGAR
ENERGY
PROTEIN
ENERGY
MSeS'23

BACTERIA USE INFORMATION FROM THEIR ENVIRONMENT TO PREDICT THE FUTURE. THE MOST RECENT INFORMATION IS THE MOST PREDICTIVE, BUT ALSO THE MOST EXPENSIVE.

© Studio Sardien

SAMENVATTING

Een fascinerende eigenschap van het leven is het vermogen van organismen om toekomstige gebeurtenissen te voorspellen. Normaal gesproken associëren we zo een complexe taak vooral met mensen, of wellicht met andere dieren. Maar in werkelijkheid vertonen zelfs bacteriën, een van de simpelste levensvormen, voorspellende capaciteiten. Dit soort eencellige organismen leven vaak in sterk variërende omgevingen, en het tot stand laten komen van de juiste reactie op een verandering kost tijd. Door veranderingen in hun omgeving te anticiperen, kunnen cellen al vooraf beginnen te reageren, wat mogelijk een evolutionair voordeel biedt. Hoe groot dit evolutionaire voordeel is hangt af van de nauwkeurigheid van de voorspelling. Maar, tot op heden is het niet duidelijk hoe nauwkeurig eencelligen de toekomst kunnen voorspellen, en waardoor die nauwkeurigheid wordt bepaald.

Voor ieder voorspellend systeem geldt dat alle informatie over de toekomst moet worden verkregen uit de huidige of vroegere omgeving. De hoeveelheid voorspellende informatie die een cel kan verkrijgen wordt daarom inherent beperkt door de hoeveelheid informatie die het uit het verleden verzamelt. In Hoofdstuk 2 onderzoeken we deze fundamentele *informatielimiet* voor twee verschillende klassen ingangssignalen. We laten zien dat, voor een Markoviaans signaal[2] , het optimale systeem simpelweg de meest recente signaalwaarde naar zijn output moet kopiëren om de informatielimiet te bereiken. Echter, als het ingangssignaal vervormd wordt door hoogfrequente ruis (snelle fluctuaties die geen nuttige informatie bevatten), dan moet het systeem niet alleen de meest recente signaalwaarde meten, maar ook eerdere waardes gebruiken om de ruis uit te middelen. Verder laten we zien dat voor een eenvoudige klasse van niet-Markoviaanse signalen, het optimale systeem zijn output moet baseren op zowel de meest recente signaalwaarde als de afgeleide daarvan. Als het doel is om meerdere eigenschappen van het signaal tegelijk te voorspellen moet het optimale systeem ook meerdere outputcomponenten gebruiken, mits de hoeveelheid beschikbare informatie uit het verleden groot genoeg is.

Welke signaaleigenschappen een biochemisch netwerk kan detecteren, hangt af van de topologie ervan. In Hoofdstuk 3 tonen we aan dat het canonieke push-pull-motief in principe de informatielimiet kan bereiken voor een Markoviaans ingangssignaal, omdat het de signaalwaarde kopieert. Tegelijkertijd hangt de hoeveelheid informatie die het netwerk kan verzamelen uit het verleden af van de fysieke middelen die de cel hiervoor beschikbaar heeft. Het netwerk gebruikt eiwitten die continu gesynthetiseerd moeten worden om verdunning door groei te voorkomen, en chemische energie om het push-pull-motief aan te drijven. Op basis van dit inzicht definiëren we een kostenfunctie voor

[2] Een Markoviaans signaal is een signaal waarvan de waarschijnlijkheid op een toekomstige waarde onafhankelijk is van het verleden, gegeven het heden. Dit betekent dat als de huidige signaalwaarde bekend is, er geen additionele informatie over de toekomst in het verleden van het signaal zit.

het totaal aan benodigde middelen van het netwerk. Het verlagen van het totaal aan beschikbare middelen onthult dat het bereiken van de informatielimiet theoretisch mogelijk is, maar ook buitengewoon kostbaar. Dit heeft twee redenen: ten eerste divergeert de chemische energie die nodig is om het netwerk aan te drijven naarmate de kopieersnelheid toeneemt; ten tweede, zelfs in een regime waarin de kosten van het aandrijven van het netwerk verwaarloosbaar zijn, kan de cel zowel de informatie uit het verleden als de voorspellende informatie vergroten door de toestanden van de receptoren—gebonden of ongebonden—te middelen over de tijd. Door zowel de voorspellende informatie als die uit het verleden tegelijk te verhogen, zorgt deze strategie ervoor dat het systeem verder van de informatielimiet verwijdert raakt. Deze observatie benadrukt dat niet alle bits aan historische informatie even voorspellend, noch even kostbaar zijn. Voor het push-pull-netwerk, en mogelijk meer in het algemeen, is de meest recente informatie uit het verleden het meest voorspellend, maar ook het duurst om te verkrijgen.

In Hoofdstuk 4 richten we ons op een belangrijk en veel bestudeerd biologisch scenario: chemotaxis in flauwe concentratiegradiënten[3]. Onze analyse toont aan dat het chemotaxisnetwerk van *Escherichia coli* de informatielimiet kan bereiken voor het voorspellen van veranderingen in de concentratie van een signaal zoals *E. coli* deze waarneemt. Maar ook hier blijkt dat het bereiken van de informatielimiet buitengewoon kostbaar is. Om de limiet te bereiken moet het netwerk namelijk een instantane afgeleide nemen. Maar, als de middelen van een netwerk dat een temporale afgeleide neemt beperkt zijn, neemt de zogenaamde gain (de versterkingsfactor) van het netwerk af met de tijdschaal waarover de afgeleide wordt gemeten. Deze tijdschaal noemen we de adaptatietijd. Om het signaal te versterken boven de onvermijdelijke biochemische ruis wanneer middelen beperkt zijn, moet de cel dus de adaptatietijd verlengen. Dit verhoogt wederom zowel de voorspellende informatie als die uit het verleden, waardoor het netwerk verder van de informatielimiet verwijdert raakt. Door de historische en voorspellende informatie van het *E. coli* chemotaxisnetwerk rechtstreeks te berekenen op basis van experimentele data wordt duidelijk dat, bij beperkte middelen, het netwerk optimaal voorspelt in de flauwste concentratiegradiënten, die ook het moeilijkst te meten zijn. Dit suggereert dat het voor *E. coli* belangrijk is om de beschikbare middelen voor voorspelling zo efficiënt mogelijk te gebruiken in flauwe gradiënten.

E. coli is niet het enige organisme dat temporele afgeleiden berekent. Het vergelijken van het huidige signaal met dat uit het recente verleden, over de adaptatietijd, is een veelgebruikte navigatiestrategie. Het is tot op heden echter onduidelijk wat de nauwkeurigheid van dergelijke metingen beperkt, en wat de optimale adaptatietijd bepaalt. In Hoofdstuk 5 breiden we het eerder ontwikkelde ‘sampling framework’ uit om het toepasbaar te maken op voorspelling, en gebruiken we dit framework om te bestuderen hoe de statistiek van het ingangssignaal en de beschikbare middelen voor het signaaltransductienetwerk het optimale ontwerp van het *E. coli* chemotaxisnetwerk bepalen. We vinden dat een optimale adaptatietijd voortkomt uit de balans tussen twee verschillende soorten fouten bij de meting van de afgeleide: de sampling fout, een statis-

[3] Chemotaxis is het gedrag van organismen waarbij ze zich verplaatsen op basis van de concentratie van bepaalde stoffen in de omgeving. In flauwe concentratiegradiënten zijn de veranderingen in de concentratie klein, en dus moeilijk te meten. Het chemotaxisnetwerk van *E. coli* is het signaaltransductienetwerk dat de cel in staat stelt zulke veranderingen waar te nemen.

tische meetfout veroorzaakt door de stochastische aard van het biochemische netwerk, en de dynamische fout, een systematische fout veroorzaakt door niet-informatieve variaties in het ingangssignaal gedurende de meting. Meer beschikbare middelen verminderen de sampling fout, waardoor een kortere adaptatietijd mogelijk wordt, wat op zijn beurt de dynamische fout vermindert. Tot slot onderzoeken we hoe de optimale adaptatietijd afhangt van de steilheid van de chemische gradiënten waarin de cel navigeert, en vinden dat de optimale adaptatietijd omgekeerd evenredig is met de steilte van de gradiënt. Dit komt doordat steilere gradiënten het signaal versterken boven de ruis en dus de sampling fout verlagen.

In alle hoofdstukken gebruiken we Gaussische ingangssignalen en bestuderen we biochemische signaalnetwerken binnen een lineaire benadering. Eerder onderzoek heeft aangetoond dat deze benadering een goede schatting geeft van de instantane *wederzijdse informatie*[4] tussen input en output van niet-Gaussische systemen. Informatie wordt echter vaak gecodeerd in de dynamiek van de volledige input- en outputtrajecten. Deze informatie kan worden gekwantificeerd door de *wederzijdse informatiesnelheid*. Er bestaat een Gaussische benadering om de wederzijdse informatiesnelheid te bepalen, maar het het exact berekenen hiervan is notoir moeilijk door de hoge dimensionaliteit van de trajectruimte. Daarom is het onbekend hoe nauwkeurig de Gaussische benadering van de informatiesnelheid is voor niet-Gaussische systemen. In Hoofdstuk 6 maken we gebruik van een recente methode die het mogelijk maakt om de wederzijdse informatiesnelheid exact te berekenen. Op basis hiervan kunnen we vervolgens de nauwkeurigheid van de Gaussische benadering evalueren. We vinden dat voor discrete lineaire reactienetwerken de Gaussische benadering slechts een ondergrens biedt voor de echte informatiesnelheid, omdat het geen onderscheid kan maken tussen afzonderlijke reacties, zoals synthese en afbraak van hetzelfde molecuul. Voor continue niet-lineaire reactienetwerken is de Gaussische benadering nauwkeurig als de gain van het netwerk voldoende klein is, zodat de input-outputrelatie van het netwerk ongeveer lineair is binnen het typische bereik van het ingangssignaal. Bovendien wordt de dynamische input-outputrelatie van het netwerk meer lineair naarmate de respons van het netwerk trager is. In dit geval kan de Gaussische benadering ook nauwkeurig zijn, maar alleen als de relevante correlatiefuncties van het niet-lineaire netwerk direct uit experimentele of simulatiedata worden geschat, en niet via een lineaire benadering van het netwerk worden berekend.

[4] Wederzijdse informatie, of mutual information in het Engels, is een grootheid uit de informatietheorie die kwantificeert in welke mate de kennis van één kansvariabele de onzekerheid in een andere kansvariabele verkleint. De meest gebruikelijke eenheid van de wederzijdse informatie is de bit. De wederzijdse informatie*snelheid* kwantificeert in welke mate de ene variabele de onzekerheid over de andere variabele verkleint *per tijdseenheid*, vaak uitgedrukt in bits per seconde. Een vertrouwd voorbeeld van een informatiesnelheid is de downloadsnelheid op onze telefoon of computer.

ABOUT THE AUTHOR

Age Tjalma is a PhD candidate in the Biochemical Networks group led by Pieter Rein ten Wolde at the research institute AMOLF. He completed his bachelor's degree in Innovation Science at Utrecht University in 2015, with a focus on life sciences. During his undergraduate studies he engaged in various pursuits, including co-founding a start-up to accelerate sustainable transport, and advising companies on energy efficiency.

Motivated to be directly involved in an emerging field of study, he began a master's degree in Bioinformatics and Systems Biology in Amsterdam in 2016. In 2017, he participated in the international synthetic biology competition iGEM, in which he and his team developed a light-fueled cell factory in the lab. To learn more about theoretical research, he subsequently started an internship at the Systems Biology lab at VU Amsterdam in 2018. There, he co-wrote a successful research proposal on bacterial growth strategies, securing a grant that allowed him to conduct both theoretical and experimental research for an additional year after completing his master's degree.

In 2020, Age began his doctoral research on optimal cellular prediction at AMOLF. This work investigates how cells can exploit past signals to obtain information about future conditions. His findings are presented in this dissertation.

LIST OF PUBLICATIONS

Appearing in this dissertation:

Age Tjalma*, Manuel Reinhardt*, Anne-Lena Moor, Christoph Zechner, and Pieter Rein ten Wolde, *Mutual Information Rate — Linear Noise Approximation and Exact Computation*, in preparation (*shared first author).

Vahe Galstyan, **Age Tjalma**, and Pieter Rein ten Wolde, *Intuitive dissection of the Gaussian information bottleneck method with an application to the problem of prediction*, in preparation.

Age Tjalma, and Pieter Rein ten Wolde, *Predicting concentration changes via discrete receptor sampling*, Physical Review Research **6**, 03049 (2024).

Jenny Poulton, **Age Tjalma**, Lotte Slim, and Pieter Rein ten Wolde, *Predicting a noisy signal: the costs and benefits of time averaging as a noise mitigation strategy*, arXiv:2307.03006 (2023), preprint.

Age Tjalma, Vahe Galstyan, Jeroen Goedhart, Lotte Slim, Nils Becker, and Pieter Rein ten Wolde, *Trade-offs between cost and information in cellular prediction*, Proceedings of the National Academy of Sciences **120**, e2303078120 (2023).

Other work:

Anne-Lena Moor, **Age Tjalma**, Manuel Reinhardt, Pieter Rein ten Wolde, and Christoph Zechner, *Reaction-Based Information Processing in Biochemical Networks*, in preparation.

Daan de Groot*, **Age Tjalma***, Frank Bruggeman, and Erik van Nimwegen, *Effective bet-hedging through growth rate dependent stability*, Proceedings of the National Academy of Sciences **120**, e2211091120 (2023) (*shared first author).

Age Tjalma, Robert Planqué, and Frank Bruggeman, *Poor sensing maximises microbial fitness when few out of many signals are sensed*, bioRxiv 800292 (2020), preprint.

ACKNOWLEDGEMENTS

These final pages of my dissertation I would like to dedicate to all the people that supported me in one way or another over the past four years. I am afraid I cannot list everyone, but know that, even if you do not find your name on these pages, I am very grateful for your support.

First and foremost, thank you Pieter Rein, for your endless enthusiasm during conversations about essentially any topic, and the many hours you have made available to discuss my work. I have learned a lot from you about science, but also about history, sailing, and countless other subjects. Most importantly, I have always felt you were sincere, and aiming to make the most of our work together.

Then Bela, thank you for being the glue of the theory groups. The digital coffee breaks during the pandemic were essential to get to know everyone, and the real coffee breaks later on were a small moment of joy in every workday. Apart from coffee, your humor and sound advice made life at AMOLF that much more fun.

Of course, I have to thank all of my colleagues. You made AMOLF feel like a second home, and I am lucky to have had such kind and generous people surrounding me. Thank you Ramon, for great conversation and being the best source for essentially all news. Thank you Alex, for your priceless knowledge ranging from the best croissants to the worst diseases. Thank you Jenny, for your openness and your fresh view on life. Thank you Mareike, for the warm welcome you provided into the group, and showing me how to be a successful PhD student. Your experience and advice had a large part in helping me endure through my first PhD year. Thank you Mike, for your knowledge on just about anything. I honestly cannot remember asking you something you didn't have an answer to. Thank you Daan (Mulder), for the many stories that came forth from your mathematician/physicist/surfer/writer/singer/construction worker life. Are there things you cannot do? Thank you Vahe, both for your kindness and for your critical mind. Whenever I got stuck, I knew I could always walk by your office for new insights. Thank you Avishek, for your warmth and enthusiasm. I greatly enjoyed our trip to Trieste, and I believe you are one of the most complete scientists I have ever met. Thank you Ariane and Anne-Sophie, for making our office a somewhat chaotic but lively place. And finally, thank you Manuel, for being able to share this PhD journey with you from day 1 to day 1461. It was reassuring to know you were in the same situation as I was, and it was inspiring to see how you dealt with any hardships along the way. You were always great company no matter where we were: in the office, exploring LA, hiking the Channel Islands, or on the climbing wall.

Een speciaal woord van dank gaat uit naar de support afdelingen van AMOLF. Jullie werk heeft mijn leven aanzienlijk makkelijker gemaakt. Specifiek wil ik Petra bedanken, dankzij jou heeft mijn onderzoek zoveel meer mensen bereikt.

De route die ik heb afgelegd voor de start van mijn PhD was allesbehalve standaard. Dat ik toch aan deze PhD ben begonnen is te danken aan een aantal mensen. Allereerst

Daan (de Groot), jij liet me zien hoe leuk theoretisch onderzoek kan zijn, van brainstormen voor het whiteboard tot op de racefiets. Ik denk nog altijd met veel plezier terug aan mijn jaar als 'jouw werknemer'—en dat zegt wat want in dat jaar brak er ook een pandemie uit. Dat ik überhaupt aan theoretisch onderzoek begonnen ben is dankzij Frank. Al tijdens de vakken die jij gaf in de master systeembiologie werkte je enthousiasme aanstekelijk. Tijdens mijn stage raakte ik verder geïnspireerd door jouw perpsectief op de biologie. Jij was ook de eerste die mij het benodigde zelfvertrouwen gaf om een PhD te overwegen, waarna dat door Daan zodanig versterkt werd dat ik het ook middenin de eerder genoemde pandemie aandurfde. Bedankt allebei.

Lieve Josine, Yuki, en Bram, wij hebben elkaar denk ik het best leren kennen tijdens iGEM. Natuurlijk samen met Thijs en Max, en onder begeleiding van Wei en Filipe. Dat halfjaar heeft mij laten zien hoe leuk het kan zijn om hard te werken aan een gezamenlijk doel, als je het maar met zulke fantastische mensen doet. Maar naast hard werken heb ik met jullie, gezamenlijk en apart, ook op hoog niveau van het leven genoten. Ieder van jullie is op je eigen manier een voorbeeld en inspiratiebron voor mij, en daar ben ik jullie heel erg dankbaar voor.

Khalid, Ruben, Marcel, Onno, Jesse, Olivier, Puck, Tijmen, Koen, Peter, Jasper, Keje, Maarten, Johan en Charlie. Aan jullie allemaal: bedankt voor de nodige afleiding van het werk. Ik heb ontzettend genoten van alle spelletjes, biertjes, koffies, diners, weekends, fietstrips, en ga zo maar door. Zonder de ontspanning die jullie verzorgden was de benodigde inspanning om mijn PhD te halen onmogelijk geweest.

Lieve familie Muffels, bedankt voor alle diners en gezelligheid. Ik voel me altijd welkom bij jullie. In het speciaal wil ik Maurice bedanken. Na publicatie van mijn eerste paper heb jij het direct opgezocht en gelezen. Jouw conclusie—dat cellen noise kunnen horen—was dan wel niet helemaal juist, het was wel een voorbeeld van jouw brede en oprechte interesse. Je was een bijzondere man, en ik ben blij je gekend te hebben.

Lieve Siets en Tjits, jullie grappen weleens dat ik de eer van ons gezin hoog moet houden door als enige een doctoraat te halen. Maar jullie zijn juist zo vaak mijn voorbeeld geweest. En terwijl jullie een paar jaar voor mij uit de paden baanden, kon ik in jullie kielzog rustig doen waar ik zin in had. Bedankt dat ik altijd bij jullie terechtkan, en voor het zijn van zulke lieve grote zussen.

Lieve pap en mam, één van de dingen die ik van jullie heb geleerd is dat een mensenleven niet planbaar is. Jullie advies was dan ook altijd om vooral nu te doen wat me nu leuk en interessant lijkt. Tegelijkertijd heb ik van jullie meegekregen om niet zomaar op te geven, en altijd mijn best te doen. Ik weet niet of dit een diep doordachte opvoedstrategie was, maar ik heb er veel aan gehad. Mijn hele studieloopbaan en de activiteiten daarnaast zijn hier een manifestatie van. Bedankt daarvoor.

Tot slot, lieve Irena, bedankt voor alles. Bedankt voor het zijn van voorbeeld en steun, van adviseur en luisterend oor, van rustpunt en vermaak. Het enigszins eenzame karakter van onderzoek, zeker tijdens een pandemie, werd altijd ruimschoots goedgemaakt door jouw gezelschap. Zonder jou had ik het niet volgehouden. Sterker nog, ik was er niet eens aan begonnen. De passie waarmee jij het leven beleeft inspireert mij, en nodigt mij uit er zelf ook voor te gaan. Wil jij de golven blijven maken? Dan surf ik met je mee.